1993

Strategies for computerized adaptive testing: golden section search, dichotomous search, and Z-score strategies

Beiling Xiao
Iowa State University

Order Number 9321228

Strategies for computerized adaptive testing: Golden section search, dichotomous search, and Z-score strategies

Xiao, Beiling, Ph.D.

Iowa State University, 1993

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

Strategies for computerized adaptive testing:

Golden section search, dichotomous search, and Z-score strategies

by

Beiling Xiao

A Dissertation Submitted to the

Graduate Faculty in Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Department: Psychology
Major: Psychology

Approved:

Signature was redacted for privacy.

In Charge of Major Work

Signature was redacted for privacy.

For the Major Department

Signature was redacted for privacy.

For the Graduate College

Iowa State University
Ames, Iowa
1993

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

Special thanks go to my major professor, Dr. Robert F. Strahan, who has graciously contributed time, guidance, expert advice, and whole-hearted support to my research.

Sincere thanks go to the members of my advisory committee: to Dr. Frederick G. Brown, who has provided valuable suggestions and references to me, and helped me to access the statistics of a SAT Verbal item pool of the College Board; to Dr. Leroy Wolins, who has provided me statistical guidance and creative ideas; to my minor professor, Dr. Yasuo Amemiya, who has given me statistical theory guidance and suggestions; and to Dr. Thomas Andre, who has contributed suggestions and computer programming guidance in my research.

Gratitude must go to Dr. Tse-Chi Hsu of University of Pittsburgh, who has reviewed one of my earlier research papers on GSSS and provided me suggestions for my future research; and to Dr. Gary Marco and Dr. Linda Cook of Educational Testing Service and Dr. Nancy K. Wright of the College Board, who have helped me to get the SAT Verbal 3-PL item parameters from recent SAT administrations.

CHAPTER 1. INTRODUCTION

Computerized adaptive testing (CAT), also called tailored testing, refers to using a computer to administer test, in which the presentation of the next test item, or the decision to stop, is adaptive. Bunderson, Inouye, & Olsen (1989) concluded that there are three categories of CAT: "*adapting item presentation*, based on item response theory parameters, particularly the item difficulty parameter; *adapting item presentation times*, based on previous response times; and *adapting the content or composition of the item*, based on previous choices" (p. 381). In the present research, CAT refers to the first category—adapting item presentation based on item response theory parameters.

In a conventional test, each individual takes the same set of items, regardless of one's ability level. In a peaked conventional test, items are chosen from a small range of difficulty levels. The peaked conventional test can distinguish people whose ability levels are suited for that difficulty range, but it can't provide accurate measurement for people whose ability levels are far away from that difficulty range. In a rectangular conventional test, item difficulties are uniformly distributed along a wide range of ability levels, and only a few items are appropriate for individuals at each ability level. The rectangular conventional test can provide almost equal but less accurate ability estimate for individuals at each ability level.

Many studies (e.g., Weiss, 1976,1982, 1983; Weiss & Kingsbury, 1984; Weiss & Vale, 1987; Lord, 1980; Roid, 1986; Stocking, 1987; Hulin, Drasgow, & Parsons, 1983, pp. 210-234; Moreno, Wetzel, McBride, & Weiss, 1984) have shown that adaptive testing procedure offers promise for improving measurement quality in many respects. A good CAT can improve both measurement quality and measurement efficiency, resulting in measurements of equal precision at all ability levels. The number of responses required in order to reach a specific level of measurement precision for an individual is substantially reduced, thus less time is required to complete the test. A general finding is that CAT can reduce test length by an average of 50%, without losing measurement quality (see, e.g., Weiss, 1979; Weiss & Vale, 1987; McBride, 1986). CAT also controls the measurement precision, in which everyone can be measured with a prespecified degree of precision. Another advantage of CAT is that it can improve standardization of the testing process and test security. Disadvantages of CAT include that CAT is more complex than conventional testing, and a CAT usually requires a large item pool that has been calibrated in advance, characterized by items with high discrimination, a rectangular distribution of difficulty, and low guessing parameters.

Strategies for Adaptive Testing: Early Work

Binet was the first person to use adaptive testing procedures (Weiss, 1982). In the Binet tests, items were presented to a child according to the child's previous responses; the examiner chose subsequent items to be administered that were most appropriate in difficulty for a child's ability as exhibited on the previous items.

Virtually each CAT strategy contains three aspects: item selection, ability

estimation, and test termination. Ability estimation includes ability estimation during the item selection process (called the current ability estimation in the present paper) and the final ability estimation. The current ability estimation and the final ability estimation can be the same ability estimation methods, or different ability estimation methods. Any ability estimation method can be used in conjunction with any item selection technique. Any test termination criterion or combination of termination criteria can be used in any CAT. Early work on adaptive testing did not always use computers, and was not based on item response theory (IRT). Early CAT strategies use pre-structured item pools. The followings are some examples.

Two-Stage tests

The two-stage test (Angoff, 1958; Betz & Weiss, 1973, 1974; Lord, 1980, pp. 128-149) consists of a short routing test, and several longer measurement tests. An individual answers the routing test first. The test is scored immediately. Based on the ability level shown on the routing test, the individual takes the measurement test on which the average difficulty level of test items is more-or-less appropriate to him or her. There are several scoring methods for the two-stage test. A simple ability estimate is the average difficulty score—the average difficulty of the items that the person answered correctly.

The two-stage test can reduce test length without degrading the measurement accuracy. However, there are several problems with the two-stage test. The main problem is its minimal adaptability; item difficulty is adapted only once during a testing procedure. Further the short length of the routing test results in routing

errors, some people may be assigned to a measurement test that is not appropriate to their ability levels. Thus the final ability estimate may not be accurate.

Flexilevel tests

A flexilevel test (Lord, 1971a, 1971b, 1980, pp. 114-127) contains an odd number of items that are ordered in difficulty. Examinees first take the item with median difficulty, and then score that item themselves. If the response is correct, the examinee proceeds to the next more difficult item; if incorrect, proceeds to the next less difficult item. This procedure is continued so as to access more or less difficult item according to one's correctness of responses. The test terminates when the examinee completes a predetermined number of items. The score on a flexilevel test is the number of correct answers (for examinees who answer the last item correctly) or the number of correct answers plus one-half point (for examinees who answer the last item incorrectly).

Lord (1971b) demonstrated that a flexilevel test can be constructed to be nearly as effective as a peaked conventional test in the ability range where the conventional test is most informative. At other ability ranges, the flexilevel test provides a better measurement than the conventional test. However, the difficulty of the items attempted by an examinee is not necessarily optimal for the examinee.

Branching tests

In a branching test, items are ordered by difficulty, and an examinee begins with an item of moderate difficulty. The general algorithm for selecting items is to proceed to a more difficult item when the previous item is answered correctly, and to proceed to a less difficult item when the previous item is answered incorrectly. There

are various branching strategies, such as pyramidal procedure (Krathwohl & Huyser, 1956; Lord, 1970), Robbins-Monro procedure (Lord, 1971c), stradaptive (stratified-adaptive) procedure (Weiss, 1973), etc. Some use a constant step size—the difference between difficulties of the item just administered and of the item to be administered next is constant. Some use shrinking-step size—the difference between difficulties of the item just administered and of the item to be administered next decreases during testing. One of the shrinking-step size procedures is the Robbins-Monro procedure (Lord, 1971c). It uses the final difficulty score as the ability estimate. When step size decreases in accordance with the conditions stated by Robbins and Monro (1951), the difficulty of the final item is a consistent estimator of the latent trait (ability) (Lord, 1971c). In his Monte Carlo study, Lord found that a Robbins-Monro test provides good measurement for a much wider range of examinee ability than does the standard test (p. 14). The Robbins-Monro procedure provides rapid convergence on ability at all levels of ability. However, it requires a large number of items to calibrate a test. An n-item Robbins-Monro test requires $2^n - 1$ items. A 20 item Robbins-Monro test requires 1048575 items, thus would be impossible to use in practical testing situations. Robbins-Monro test is also subjected to guessing effects.

In a stratified adaptive computerized ability test (Weiss, 1973), items are stratified by difficulty level or organized into a set of scaled peaked tests. By using certain branching procedures, the examinee is branched to the region of the item pool that provides the maximum information about his or her ability level. It can use various kinds of average difficulty scores, or the highest item difficulty score (the difficulty of the most difficult item the examinee answered correctly), as the examinee's ability estimate.

IRT-Based Computerized Adaptive Tests

Overview of the IRT

IRT "is a family of theories and models which have been developed to measure hypothetical construct, or constructs, from individuals' reactions to particular stimuli" (Samejima, 1983, p. 159). It provides a convenient way to assess an individual's ability on the common metric, even when different persons are tested using different sets of items. The item difficulty also shares the same metric with the ability scale.

IRT models specify the relationship between the probability of the observed responses of an individual to a test item and the individual's ability level. Models include the normal ogive model and the one-, two-, and three-parameter logistic models (Lord, 1952; Birnbaum, 1968). The three-parameter logistic (3-PL) model is (see Lord, 1980, Equation 2-1; Hulin, et al., 1983, Equation 2.4.4):

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + \exp[-Da_i(\theta - b_i)]} \tag{1.1}$$

where θ is the latent trait (ability) value; a_i reflects the steepness of the item response function (i.e., the item response curve (ICC)) at its inflection point, and is an index of item discrimination; b_i is an index of item difficulty and corresponds to the value of θ at the inflection point of the ICC; c_i is the lower asymptote of the ICC (also called the pseudo-guessing parameter), and corresponds to the probability of a keyed response among examinees with very low levels of θ; D is a scaling constant (in the present research, $D = 1.702$ or 1.7. Using these values of D, the logistic $1/[1 + \exp(-Dx)]$ is approximately equal to the probit), and $P_i(\theta)$ is the probability of a keyed response to item i for people whose ability level is θ.

When $c_i = 0$, Equation 1.1 becomes:

$$P_i(\theta) = \frac{1}{1 + \exp[-Da_i(\theta - b_i)]} \tag{1.2}$$

Equation 1.2 is the two-parameter logistic (2-PL) model.

When $c_i = 0$ and $a_i = 1$, Equation 1.1 becomes:

$$P_i(\theta) = \frac{1}{1 + \exp[-D(\theta - b_i)]} \tag{1.3}$$

Equation 1.3 is the one-parameter logistic (1-PL) model. The 1-PL and the 2-PL models thus are special cases of the 3-PL model.

When c_i in Equation 1.1 is fixed, that is , the c_i parameter is a constant, and $0 < c < 1$, Equation 1.1 is called the modified 3-PL model (Green & Bock, 1984). The a_i parameter is also a constant in the modified 3-PL item pool of the present research.

IRT theory has several strong assumptions. The major assumptions are: (a) local independence—responses to any two items are uncorrelated in a homogeneous subpopulation at a particular ability level; and (b) a specified shape for the ICC. Based on IRT, several CAT strategies have been developed. There are currently two primary IRT-based item selection methods: the maximum item information selection (Lord, 1980) and the Bayesian item selection (Owen, 1969, 1975). There are also two primary IRT-based ability estimation (i.e., scoring) methods: the maximum likelihood estimation and the Bayesian scoring approaches (Owen, 1969, 1975; Bock & Mislevy, 1982). The maximum information item selection is usually used in conjunction with the maximum likelihood estimation, while the Bayesian item selection is usually used in conjunction with the Bayesian scoring. Combining these two item

selection methods and scoring approaches in a reversed way is also possible. These CAT strategies usually use a prespecified degree of precision as termination criterion.

The maximum likelihood estimate of ability

If the item parameters are known from pretesting, the maximum likelihood estimate (MLE) of ability can be obtained by solving for θ in the following equation (see Lord, 1980, Equation 5-19):

$$\sum_{i=1}^{n} \frac{P_i\prime(\theta)}{Q_i(\theta)} = \sum_{i=1}^{n} W_i(\theta)\mu_i \tag{1.4}$$

where n is the number of items, $Q_i(\theta) = 1 - P_i(\theta)$, $P_i\prime(\theta)$ is the derivative of $P_i(\theta)$ with respect to θ, $\mu = 1$ when the item response is a keyed answer, and $\mu = 0$ when the item response is not a keyed answer, and $W_i(\theta)$ is the locally best weight (Birnbaum, 1968) or the optimal scoring weight (Lord, 1980), and

$$W_i(\theta) = \frac{P_i\prime(\theta)}{P_i(\theta)Q_i(\theta)} \tag{1.5}$$

For the 3-PL model the locally best weight is:

$$W_i(\theta) = \frac{Da_i}{1 + c_i \exp[-Da_i(\theta - b_i)]} \tag{1.6}$$

For the 2-PL model the locally best weight is:

$$W_i(\theta) = Da_i \tag{1.7}$$

For the 1-PL model the locally best weight is:

$$W_i = 1 \tag{1.8}$$

When the item responses μ_i are all equal to zero or all equal to one, Equation 1.4 has no root. When n is small, Equation 1.4 may have more than one root (Samejima,

1973). Lord (1980, p. 59) indicated that if n is large enough, the uniqueness of the root of the MLE equation is guaranteed. He also indicated that with item number $n \geq 20$, multiple roots had not been found in practical work. However, Yen, Burket & Sykes (1991) examined the response vectors for large groups of real examinees in fourteen multiple-choice achievement tests with 20 to 50 items, found that from 0.0 to 3.0% of them had response vectors with multiple maxima. The unique root of Equation 1.4, is a consistent estimate of θ, and converges to the true parameter value, θ. The MLE is also asymptotically normal and efficient under certain regularity conditions (Hulin et al., 1983, p. 48). The sample distribution of the MLE becomes normally distributed as sample size increases. In large samples the sampling variance of the MLE reaches a theoretical lower bound.

The asymptotic sampling variance of the MLE is (see Lord, 1980, Equation 5-5):

$$Var(\hat{\theta}|\theta) = \frac{1}{\sum_{i=1}^{n} \frac{P_i{}'^2(\theta)}{P_i(\theta)Q_i(\theta)}} \tag{1.9}$$

The (asymptotic) information function, $I\{\theta\}$, of the MLE of ability is the reciprocal of the asymptotic variance (Lord, 1980, Equation 5-6):

$$I\{\theta\} \equiv I\{\theta, \hat{\theta}\} = \sum_{i=1}^{n} \frac{P_i{}'^2(\theta)}{P_i(\theta)Q_i(\theta)} \tag{1.10}$$

and is the maximum information provided by a test. The information provided by each item is (Lord, 1980, Equation 5-9):

$$I\{\theta, \mu\} = \frac{P_i{}'^2(\theta)}{P_i(\theta)Q_i(\theta)} \tag{1.11}$$

The item information contributes independently to the test information function.

The item information for the 3-PL model is:

$$I\{\theta\} = \frac{D^2 a_i^2 (1 - c_i) \exp[-Da_i(\theta - b_i)]}{\{1 + c_i \exp[-Da_i(\theta - b_i)]\}\{1 + \exp[-Da_i(\theta - b_i)]\}^2} \tag{1.12}$$

The item information is a bell shaped curve as a function of θ. For the 3-PL model an item gives maximum information at ability level $\theta = \theta_i$ (Lord, 1980, Equation 10-4) where

$$\theta_i = b_i + \frac{1}{Da_i} \ln \frac{1 + \sqrt{1 + 8c_i}}{2} \tag{1.13}$$

When

$$b_i = \theta_i - \frac{1}{Da_i} \ln \frac{1 + \sqrt{1 + 8c_i}}{2} \tag{1.14}$$

an item with difficulty b_i can maximize information at ability level $\theta = \theta_i$ (see Birnbaum, 1968, Equations 20.4.21-24; Lord, 1980, Equation 11-26).

Maximum information item selection

The maximum information item selection selects, from the item pool, the most "informative" item for an individual to respond to, based on the current ability estimate. For a particular level of ability θ, information of each item in the item pool at ability θ can be calculated. The item that has the maximum information will be selected. If there is no prior information about an examinee's ability range, the item selection process usually starts by assuming the individual's ability is equal to the population mean, and presenting him/her with the most informative item according to this assumed ability level. After the individual makes at least one correct and at least one incorrect responses, the MLE is calculated. Based on this MLE, the most informative new item is chosen to be presented. The test terminates when certain predetermined criteria are achieved (such as a certain level of accuracy is achieved, which can be easily inferred by the test information function), or a certain number of items have been presented.

The distribution of the MLE of ability θ is approximately (asymptotically) the normal distribution with mean θ, and variance $1/I\{\theta\}$. It is asymptotically efficient (Birnbaum, 1968, p. 457; Lord, 1980, pp. 70-71).

Many CAT studies (e.g., Stocking, 1987; Hulin et al., 1983, pp. 210-234; Weiss, 1982; Green, 1983) have shown that CAT using the maximum information item selection, in conjunction with the MLE scoring, can provide very accurate ability estimates along the ability continuum. It is more efficient than the Owen's Bayesian strategies (see Weiss, 1982). However, when guessing exists, lower ability examinees can easily get higher ability estimates based upon the response patterns that contain mainly successful item responses, in the earlier stage of CAT. Substantial number of items should be administered to recover the effect of lucky guessing. In addition, if an examinee answered all the items right or all the items wrong, the MLE cannot be determined. There are also some unusual response patterns that the maximum likelihood estimate procedure fails to converge (Weiss, 1982). In the case where there is no MLE solution, arbitrary θ estimate is assigned, or arbitrary predetermined branching rules are used during the item selection process. To select the most "informative" item for the current ability estimate, from the entire item pool, the information of every item that has not been administered needs to be calculated. The computation burden is heaver. In practice, an *info table* (Thissen & Mislevy, 1990) is used to reduce the computation burden. The info table contains lists of items, ordered by the amount of information they can provide at various levels of ability. Since the current ability must be approximated into those tabulated ability values, some measurement efficiency may be lost. Computation power is not a problem for the 386 PC with a math co-processor and 486 PC. For those machines with high

computation power, no info table is necessary for applying the maximum information item selection method.

Bayesian ability estimation and item selection

The general procedure used in Owen's Bayesian CAT (see Owen, 1969, 1975; Jensema, 1974) begins with the specification of a prior distribution for θ; that is, the mean and the variance are specified. If there is not any prior information about a particular examinee, it usually assumes that θ is normally distributed with mean 0 and variance 1. An item is selected from the calibrated item pool that will most reduce the uncertainty of the examinee's ability estimate when administered; that is, most reduces the standard error of the ability estimate. After this item is presented, the prior ability estimate and the information from the response to this item are combined by means of Bayes' theorem to obtain a posterior ability estimate and the standard error of the ability estimate. The posterior ability estimate then becomes the prior ability estimate for the next stage. A new item is then chosen from the item pool that can most reduce the uncertainty of the ability estimate, and Bayes' theorem is used to combine the information obtained by administering the present item and the prior ability estimate to obtain an updated posterior ability estimate. The standard deviation of the new posterior ability estimate is also estimated. This iterative process continues until the standard error of the ability estimate is sufficiently small, or until a predetermined number of items are presented. The final posterior ability estimate is the examinee's ability estimate using the Bayesian scoring approach.

The Bayesian-based estimation is closely related to the likelihood-based estimation (Sympson, 1977; Lord, 1980; Weiss, 1982). The MLE is a special case of

the Bayesian estimate when the Bayesian prior distribution for θ is uniform. The Bayesian posterior variance is related to the variance of the likelihood function, which is inversely related to the information of a particular response pattern. Weiss (1982) indicated that approximately 85% of the items selected by the two procedures were the same. The Owen's Bayesian adaptive test strategies can achieve accurate results when the prior θ estimate is accurate, and is more efficient than most other CAT strategies (see Vale, 1975). Unlike the MLE, which sometimes has no solution or has multiple solutions, the Bayesian scoring method always gets a unique ability estimate. The computation for selection of the next item, such that administration of that item would maximally reduce the posterior variance, is not complex. However, the relationship between the Owen's Bayesian estimated score and the latent ability is nonlinear in the lower ability range (McBride, 1975; Weiss & Mcbride, 1984). When the prior θ estimate is not accurate, the estimated results are severely biased (Weiss & McBride, 1984). In Owen's Bayesian procedure, the approximate estimate of ability varies as a function of the item presentation order (Thissen & Mislevy, 1990). IRT estimates of ability should not depend on the item presentation order. Due to these problems, and the development of the microcomputer computation power, Owen's Bayesian item selection method is much less widely used.

Bock and Mislevy (1982) described the adaptive EAP (estimated a posteriori) estimation of ability in CAT. It uses numerical method to approximate the mean and variance of the Bayesian posterior distribution and is shown to have good properties for CAT. The EAP estimator has minimum mean square error over the population of ability for which the distribution of ability is specified by the prior. The EAP estimator is not affected by the item presentation order. The EAP estimator is

biased whenever a finite sample of items is used. The bias of the EAP estimate is minor for most of the population (within ± 2 standard deviations).

Other item selection methods

Though the maximum information item selection and the Bayesian item selection methods can choose from the item pool the item that is the "most informative" or that can provide the "least expected posterior standard error", they face problems, such as item exposure rate. Items with higher a values are heavily used. Items with low a values may be never used. The following item selection methods are less efficient, compared to the maximum information item selection, but they choose items more evenly from the item pool.

Match m_i to $\hat{\theta}$ item selection. The match m_i to $\hat{\theta}$ item selection (see Hulin et al., 1983) selects from the items not yet administered the item whose m_i is closest to the current ability estimate $\hat{\theta}$, to administer, where

$$m_i = b_i + \frac{1}{Da_i} \ln \frac{1 + \sqrt{1 + 8c_i}}{2} \tag{1.15}$$

From Equations 1.13 and 1.14 one can infer that item whose difficulty is b_i can provide maximum information at ability level m_i. In Equation 1.15, if $c_i = 0$, then $m_i = b_i$. That is, if $c_i = 0$, the match m_i to $\hat{\theta}$ item selection selects item whose difficulty value b_i is closest to $\hat{\theta}$.

The match m_i to $\hat{\theta}$ item selection is different from the maximum information item selection. In the former, each item's m_i value is calculated, the item that has the smallest difference between m_i and $\hat{\theta}$ is selected. In the later, each item's information at $\hat{\theta}$ is calculated, the item that has the maximum information is selected. In the

1-PL item pool, where $a_i = 1$, $c_i = 0$ for all items, or the modified 3-PL item pool, where a_i and c_i are constants for all items, the item selected by the match m_i to $\hat{\theta}$ item selection and the item selected by the maximum information item selection are identical.

Quasi-match m_i to $\hat{\theta}$ item selection. The quasi-match m_i to $\hat{\theta}$ item selection (Xiao, 1989, 1990) is a robust and simple item selection method.

A constant d is calculated for each item pool:

$$d = \frac{1}{Da} \ln \frac{1 + \sqrt{1 + 8c}}{2} \tag{1.16}$$

where a and c are the means of a_i and c_i of the whole item pool, respectively. Clearly, $d = 0$ when c_i of each item is 0. That is the case in the 1-PL or the 2-PL model.

Define $m_i = b_i + d$. If the current ability estimate is a point estimate (e.g., MLE), the quasi-match m_i to $\hat{\theta}$ item selection selects the item not yet administered whose m_i value is closest to $\hat{\theta}$. If the current ability estimate is an interval estimate, it could be smaller or greater than the MLE. In this case the quasi-match m_i to $\hat{\theta}$ item selection selects the item not yet administered whose m_i value is closest to $\hat{\theta}$, in the direction where the MLE lies; that is, whose m_i value is next smaller (or next greater) to the $\hat{\theta}$ value, depending upon whether the obtained score x is smaller (or greater) than the expected score at $\hat{\theta}$.

In the 1-PL or the modified 3-PL item pools, where the as and cs are constants, during the item selection process, when the current ability estimation is a point estimate, the item selected by the quasi-match m_i to $\hat{\theta}$ item selection is the same as the item selected by the match m_i to $\hat{\theta}$ item selection or the maximum information item selection.

Earlier CAT strategies are not based on IRT, are easier to develop, but are found much less efficient than IRT-based CAT strategies (see, e.g., Weiss, 1982; Vale, 1975). CAT strategies that are based on IRT models usually need a very large item pool, which meets assumptions of IRT models. To estimate item parameters precisely, large sample size is needed. Though IRT-based CAT has substantial advantages, as previously mentioned, and computer technology is highly developed today, CAT is still not a widely-used procedure in measurement areas. The difficulty in developing an appropriate item pool may be one reason. The complexity of IRT-based CAT strategies may be another reason. More simple, more robust CAT strategies are needed in order to enable the wide use of CAT.

In this paper, a family of CAT strategies—golden section search strategies (GSSS), dichotomous search strategies (DSS), and two versions of Z-score strategies (ZSS)—are introduced. Those strategies have several robust features. They are less subjected to the guessing effects and the inaccuracy of the item parameters. They can provide almost unbiased and accurate ability estimates over a wide range of ability levels, provide a convenient frame to use various item selection methods and scoring methods, and are computationally efficient. In GSSS, DSS, and one version of ZSS, a hypothesis testing is involved to determine the current ability estimate. The current ability estimates are interval estimates. They usually provide a more conservative current ability estimate than the MLE of ability. That is, the current ability estimate tends to be the same as the previous ability estimate, unless it statistically differs from the previous estimate. In another version of ZSS, no hypothesis testing is conducted to determine the current ability estimate. The Z-score estimate is used as the current ability estimate. Thus the current ability estimate is a point estimate. In the MLE

of ability, if an examinee responds to all items correctly or all items incorrectly, or examinee has an unusual response pattern, the MLE has no solution. Thus an arbitrary ability estimate must be assigned. The Z-score estimate can usually be a reasonable ability estimate in those situations.

Dichotomous search strategies for CAT

Xiao (1990) proposed dichotomous search strategies (DSS) for CAT. DSS for CAT is adapted from an optimization search method—dichotomous search. Dichotomous search, or dichotomizing search, is a procedure for searching for a point in an interval, in which, at each step, the interval is divided into two halves, one half being then discarded if it can be logically shown that the point could not be in that half. (see, e.g. , Hua, 1981; Adby & Dempster, 1974; Gottfried & Weisman, 1973; Parker, 1984). The process of DSS is as follows:

Testing points. Testing points are the midpoints of successive search regions. The original search region covers all item difficulty levels in the item pool, or all possible ability levels that an examinee's ability estimate would be, whichever is less. The upper or the lower half of the search region is discarded when the search process continues. The ratio of the sizes of the next search region to the previous search region is .5. The scale for the testing points and the search region is the same as the θ scale.

Hypothesis testing and the current ability estimation. After each item is administered, compute the expected score and its variance at each successive testing point θ. Compute the obtained score of the examinee at each successive testing point

θ. Test scores usually have the following form (see Birnbaum, 1968, Equation 17.7.1):

$$x = \sum_{i=1}^{n} W_i \mu_i \tag{1.17}$$

where W_i is the specified numerical weight, and $\mu_i = 1$ when the item response is a keyed answer, and $\mu_i = 0$ when the item response is not a keyed answer, n is the number of items.

The mean of x is (see Birnbaum, 1968, Equation 17.7.2):

$$\mu_{x|\theta} = \sum_{i=1}^{n} W_i P_i(\theta) \tag{1.18}$$

The variance of x is (Birnbaum, 1968, Equation 17.7.3):

$$\sigma^2_{x|\theta} = \sum_{i=1}^{n} W_i^2 P_i(\theta) Q_i(\theta) \tag{1.19}$$

where $P_i(\theta)$ is the item response function of item i, and $Q_i(\theta) = 1 - P_i(\theta)$.

After each item is administered, the hypothesis testing is conducted. Compare the obtained test score x (at θ) with the expected score $\mu_{x|\theta}$ at the testing point of the original search region. If x is within the range of $(\mu_{x|\theta} \pm Z_\alpha \sigma_{x|\theta})$, the examinee's current ability estimate is assumed to be equal to that testing point θ. If x is larger than (or smaller than) $\mu_{x|\theta}$ plus (or minus) $Z_\alpha \sigma_{x|\theta}$ (Z_α is a prespecified numerical weight. In the present research, it is called the SD weight), the examinee's ability is assumed higher (or lower) than the testing point θ. The lower (or upper) half of the search region is discarded. Hypothesis testing using updated testing point θ continues, until x is within a confidence interval of the expected score at a testing point, or until the size of the search region is smaller than a prespecified small value (in that case the last testing point will be the current ability estimate). The scale for the obtained scores and the expected scores differs from the scale of θ.

Test score weight. The numerical weight of test score in Equation 1.17 or Equation 1.18 can be as simple as unit weight (that is 1). If the unit weight is used, the test score is a number-right score. If the weight for each item is chosen by Equation 1.5, that is the locally best weight, the test score can provide ability estimate θ with optimal or nearly optimal precision. If we substitute the locally best weight into the following equation:

$$\sum_{i=1}^{n} W_i \mu_i = \sum_{i=1}^{n} W_i P_i(\theta) \tag{1.20}$$

it becomes

$$\sum_{i=1}^{N} \frac{P_i\prime(\theta)}{P_i(\theta)Q_i(\theta)} \mu_i = \sum_{i=1}^{n} \frac{P_i\prime(\theta)}{Q_i(\theta)} \tag{1.21}$$

Equation 1.21 is a maximum likelihood equation (see Equation 1.4). An examinee's MLE of ability θ can be obtained by solving Equation 1.21, if the response pattern of the examinee is known. The left side of Equation 1.20 is the examinee's obtained score, the right side of Equation 1.20 is the expected score at ability θ. If the locally best weight is used in DSS, the current ability estimate is within a confidence interval of θ, that is, the confidence interval of the MLE $= \theta$.

Lower than chance score adjustment. Equation 1.1 has a pseudo-guessing parameter c_i. Whenever an examinee's obtained score x is less than a portion of score assumed from random guessing

$$x_C = \sum_{i=1}^{n} W_i c_i \tag{1.22}$$

the obtained score x is adjusted into x_C. This adjustment is only applied during the current ability estimation process. In the final ability estimation, no such adjustment is made.

Item selection. Any item selection methods can be used in conjunction with DSS. Item selection methods include the quasi-match m_i to $\hat{\theta}$ item selection, the match m_i to $\hat{\theta}$ item selection, and the maximum information item selection.

Test termination criterion. Both of the two primary termination criteria discussed previously could be used in DSS CAT—a prespecified degree of precision, or certain number of items. The index used in the former criterion could be the test information (see Lord, 1980, pp. 65-80) at the current ability estimate, or the response pattern information at the current ability estimate (see Samejima, 1973, 1983, pp. 159-165).

The final ability estimation. The MLE scoring method could be used in DSS to obtain the final ability estimate. A lot of methods could be applied to solve the MLE solution. One of them is the modified Newton-Raphson method (Hamming, 1973, pp. 68-72). Another simplest method is to apply dichotomous search to solve the MLE of ability.

Golden section search strategies for CAT

Golden section search method. Xiao (1989) proposed GSSS. GSSS is adapted from one of the optimization search methods—golden section search (see e.g., Hua, 1981; Wagner, 1975, Chapter 14; Adby & Dempster, 1974; Walsh, 1975, pp. 91-93). The golden section search method can search for the maximum or the minimum of a unimodal function. The function is not necessarily continuous nor necessarily defined by a mathematical expression. The size of the search region which contains the maximum (or minimum) is reduced, at each iteration, by the golden section

ratio t:

$$t = \frac{\sqrt{5}-1}{2} \approx .618 \tag{1.23}$$

and, of any three successive search regions in golden section search, the size of the second search region plus the size of the third search region is equal to the size of the first search region. There are two possible cutting points in each search region. For search region $[a, b]$ $(a < b)$, the two cutting points c_1, c_2 are:

$$c_1 = b - t(b-a) \tag{1.24}$$

$$c_2 = a + t(b-a) \tag{1.25}$$

In the search process, the search region [a, b] is reduced to $[a, c_2]$, (or $[c_1, b]$), by discarding the upper portion (or the lower portion) of [a, b], from the cutting point c_2 (or c_1), if it can be logically shown that the point being searched is not in that portion. Two new cutting points will be determined in the new search region. Successive search region will reduce by a ratio t.

The GSSS is the same as DSS in all respects, except for the ratio of successive search regions. As in DSS, each midpoint of each search region in GSSS is also a testing point. Unlike DSS, there are two possible cutting points in each search region in GSSS (in DSS, there is only one cutting point in each search region, that is the midpoint of each search region). In GSSS, search region reduces by a ratio t. Similar to DSS, each testing point is at the midpoint of each search region. The hypothesis testing is conducted after each item is administered. Compare the obtained test score with the expected score at the testing point of the original search region. If an examinee's obtained score is within a confidence interval of the expected score at a testing point, the examinee's current ability estimate is assumed to be equal to that

of the testing point. Otherwise, the upper (or lower) portion of the search region, from the upper (or lower) cutting point, is discarded and the process is continued until the examinee's current ability estimate is determined.

Z-score strategies for CAT

Two versions of ZSS are proposed. One version of ZSS is similar to GSSS and DSS in applying statistical hypothesis testing in determining the current ability estimate. It is called ZSS (using SD weight). Another version of ZSS does not apply statistical hypothesis testing in determine the current ability estimate. It is called ZSS (no SD weight).

ZSS (using SD weight) is similar to DSS and GSSS in that a confidence interval is involved in determining an examinee's current ability estimate. Thus a SD weight is needed to calculate the confidence interval. In DSS and GSSS, there are systematic search regions shrunk by a ratio. Each testing point is at the midpoint of each search region. After each item is administered, statistical hypothesis testing is conducted at each of the successive testing points, to determine the current ability of the examinee. In ZSS, there is no search region. The first testing point is prespecified, which is equal to an examinee's pre-ability estimate. The successive testing point is determined by the Z-score estimate described below.

After each item is administered, an examinee's obtained score is compared with the expected score at a testing point θ, which is the same as the previous ability estimate. If the obtained score is within a confidence interval of the expected score at that testing point, the examinee's current ability estimate is assumed to be the same as the previous ability estimate. Next item to be presented will be based on

that current ability estimate. On the other hand, if the obtained score exceeds the confidence interval of the expected score at that testing point, the next testing point θ_1 will be defined by the following equation:

$$\theta_1 = \theta + Z_{x|\theta}\frac{1}{\sqrt{I\{\theta\}}} \tag{1.26}$$

where

$$Z_{x|\theta} = \frac{x - \mu_{x|\theta}}{\sigma_{x|\theta}}$$

is the Z score of the obtained score x with respect to the expected score $\mu_{x|\theta}$, $I\{\theta\}$ is the information at θ (see Equation 1.10), and it is the reciprocal of the asymptotic variance of the estimates of θ. In the present research, θ_1 is called the Z score estimate of ability for examinee whose previous ability estimate is θ. The obtained score x and ability θ are the same thing expressed on different scales of measurement. A monotonic transformation of scores should not change the relative size of the confidence interval of θ (see Lord, 1980, pp. 78-80).

Substitute the locally best weight in Equation 1.19, get

$$\sigma^2_{x|\theta} = \sum_{i=1}^{n} \frac{P_i'(\theta)^2}{P_i(\theta)Q_i(\theta)} \tag{1.27}$$

which is the same as $I\{\theta\}$ (see Equation 1.10).

Substitute $I\{\theta\} = \sigma^2_{x|\theta}$ into Equation 1.26, get

$$\theta_1 = \theta + \frac{x - \mu_{x|\theta}}{\sigma^2_{x|\theta}} \tag{1.28}$$

The hypothesis testing described above is applied to the new testing point θ_1, to test whether the obtained score is within a confidence interval of the expected score at θ_1. If the obtained score exceeds the confidence interval at θ_1, a new testing point is

calculated, and hypothesis testing will be conducted at that new testing point, until a current ability estimate is determined. The next testing item is selected according to certain item selection method. In ZSS (using SD weight), if the current ability estimate is bigger (or smaller) than a prespecified ability range, the current ability estimate is assigned to be the higher (or the lower) bound of that range. Usually that range is assigned equal to the item difficulty range of the item pool used. Other aspects other than testing point allocation are the same as in DSS and GSSS.

In GSSS, DSS, and ZSS (using SD weight), statistical hypothesis testing is involved in the process of the current ability estimation, and a confidence interval is needed in determining whether the obtained score is equal to the expected score at a testing point. The current ability estimate is an interval estimate, not a point estimate.

Another version of ZSS does not apply statistical hypothesis testing in determining the current ability estimate. No confidence interval is used in determining the current ability estimate. Thus no SD weight is needed. In ZSS (no SD weight), after each item is administered, a Z-score estimate is calculated according to Equation 1.26 (or Equation 1.28, if the locally best weight is used). The current ability estimate is the Z-score estimate with respect to the previous ability estimate. Next item to be administered is chosen based on the current ability estimate. In the current ability estimation process, if an examinee's obtained score x is less than a portion of score assumed from random guessing, the lower than chance score adjustment is applied. The MLE of ability is used as the final ability estimate. If the number of items is not small, the Z-score estimate can be used as the final ability estimate. As the number of items increases, the Z-score estimate tends to approach the MLE, provided that

the information at the previous estimate is high. There is no confidence interval associated with the current ability estimate. The current ability estimate is a point estimate in ZSS (no SD weight). In ZSS (no SD weight), if the current ability estimate is bigger (or smaller) than a prespecified ability range, the current ability estimate is assigned to be the higher (or the lower) bound of that range. Usually that range is assigned equal to the item difficulty range of the item pool used. Using a moderate SD weight, ZSS (using SD weight) provides almost the same measurement results as that provided by ZSS (no SD weight) in most item pools.

The above description about DSS, GSSS, and ZSS showed their robust nature and their potential to measure effectively. Xiao's exploration study (1989) showed that GSSS is more accurate and more efficient than the peaked conventional test and the rectangular conventional test, is as accurate and efficient as maximum likelihood estimate strategies (MLES) when there is no guessing, and could provide a more precise ability estimate than does MLES whenever guessing exists. Xiao (1990) found that both DSS and GSSS were more efficient and more accurate than MLES whenever guessing exists. DSS measured slightly better in the extremely low or extremely high ability levels than did GSSS, while GSSS measured slightly better than did DSS in the middle range of ability levels. Because these previous studies of Xiao used only low power computers (APPLE IIe, APPLE IIe Plus, or ZENITH 150), only the simplest item selection method—the quasi-match m_i to $\hat{\theta}$ item selection—was applied, in the 1-PL and the modified 3-PL item pools. Arbitrary SD weights were used in the previous studies. Thus, arbitrary sizes of the confidence interval were used in determining the current ability estimate.

The present Monte Carlo research attempted to determine the appropriate size

of the confidence interval to use in GSSS, DSS, and ZSS during the current ability estimation process, to compare the measurement precision of the two versions of ZSS and to compare the efficiency and accuracy of DSS, GSSS, and ZSS with MLES in different item pools, using different item selection methods. DSS is similar to GSSS in nature. They only differ in the ratios of their successive search regions. ZSS (using SD weight) is similar to DSS and GSSS in using a confidence interval to determine the current ability estimate. ZSS (no SD weight) is similar to ZSS (using SD weight) in computing the Z-score estimate. Four item pools were used in the present research: (a) the 1-PL item pool; (b) the modified 3-PL item pool with as and cs fixed; (c) the hypothetical 3-PL item pool; and (d) the SAT Verbal 3-PL item pool. Three kinds of computers were used: (a) ZENITH 150 IBM compatible microcomputers; (b) ZENITH 386/20 IBM compatible microcomputers; (c) Digital UNIX DEC Station 3100. Two kinds of examinee samples were involved: (a) 2500 computer simulated examinees grouped into 25 ability levels from -3 to 3, in interval 0.25; (b) 1000 computer simulated examinees whose ability levels were normally distributed with mean 0 and variance 1.

The D constant in the logistical models was 1.702 in programs running on the IBM compatible microcomputers, was 1.7 in programs running on the Digital UNIX workstations.

In Study One, Monte Carlo studies were conducted to compare the measurement precision of three CAT strategies—GSSS, DSS, and ZSS using different SD weight, in the 1-PL and the modified 3-PL item pools. In Study Two, Monte Carlo studies were conducted to compare the measurement precision of two item selection methods: the quasi-match m_i to $\hat{\theta}$, and the maximum information item selections for GSSS, DSS,

and ZSS; and to compare the measurement precision of the two versions of ZSS—ZSS (using SD weight), and ZSS (no SD weight), in the 1-PL and the modified 3-PL item pools. In Studies Three, Four, and Five, the measurement accuracy and efficiency for GSSS, DSS, ZSS, and MLES were compared. In Study Three, Monte Carlo studies were conducted to compare the measurement accuracy and efficiency of GSSS, DSS, ZSS, and MLES, using the maximum information item selection, in the 1-PL and the modified 3-PL item pools. In Study Four, Monte Carlo studies were conducted to compare the measurement accuracy and efficiency of GSSS, DSS, ZSS, and MLES, using three different item selection methods—the quasi-match m_i to $\hat{\theta}$ item selection, the match m_i to $\hat{\theta}$ item selection, and the maximum information item selection, in the hypothetical 3-PL item pool. Computational efficiency for the four CAT strategies was also checked. In Study Five, Monte Carlo studies were conducted to compare the accuracy and efficiency of GSSS, DSS, ZSS, and MLES, assuming the 3-PL and the 1-PL models, in a SAT Verbal 3-PL item pool. In Studies One, Two, Three, and Four, computer simulated examinees whose true ability levels were grouped into 25 levels were used. In Study Five, computer simulated examinees whose true ability levels were normally distributed were used.

The measurement accuracy was evaluated by the bias, absolute errors (ABE), and mean squared errors (MSE). The measurement efficiency was evaluated by test information. Virtually it is impossible to differentiate the measurement accuracy from the measurement efficiency. The classification of the measurement precision indices into two categories—accuracy and efficiency—is mainly for illustration purpose. It is based on the common sense that the accuracy is the degree of how close an estimate is to the true value, while the efficiency is an index of the maximum measurement

precision a measure can achieve.

To compare the computational efficiency of GSSS, DSS, ZSS, and MLES, executing times of some of the Monte Carlo studies were recorded.

The bias is

$$B(\theta) = \frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta) \tag{1.29}$$

The ABE is

$$ABE(\theta) = |\hat{\theta}_i - \theta| \tag{1.30}$$

The MSE is

$$MSE(\theta) = \frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta)^2 \tag{1.31}$$

The test information $I\{\theta\}$ is defined by Equation 1.10.

It is expected that GSSS, DSS, and ZSS will provide a more precise ability estimate than does MLES and will be more computationally efficient than MLES.

CHAPTER 2. STUDY ONE: APPROPRIATE *SD* WEIGHTS FOR GSSS, DSS, AND ZSS

Design

In GSSS, DSS, and ZSS (using *SD* weight), a statistical hypothesis testing is conducted to determine whether the obtained score is within a confidence interval of the expected score at a testing point. A current ability estimate is determined based on the hypothesis testing results. Monte Carlo studies were conducted to find the appropriate size of confidence interval in determining the current ability estimates for the three CAT strategies—GSSS, DSS, and ZSS, in the 1-PL and the modified 3-PL models, on two kinds of computers—ZENITH 150 IBM compatible microcomputers (with or without a hard disk drive) and Digital UNIX DEC Station 3100. The measurement precision of the ability estimates using different *SD* weights were compared. Each CAT contained 20 items. Four measurement precision indices—bias, absolute errors, MSE, and test information—were calculated. All CAT was conducted using simulated examinees. The programs running on the IBM compatible microcomputers were developed by the author using the GWBASIC language (not compiled). The programs running on the Digital UNIX workstations were written in C language (compiled) by the author.

Method

Simulees

A simulee was a computer generated examinee with a true ability value θ. On the IBM compatible microcomputers, there were 2500 simulees for each of the three CAT strategies—GSSS, DSS, and ZSS using each of the four SD weights—0.4, 0.7, 1.0, and 1.3, in the 1-PL and the modified 3-PL item pools, respectively, with 100 as a group at each of 25 ability levels equally spaced in [-3, 3], in interval of 0.25. On the UNIX workstations, there were the same 2500 simulees for each of the three CAT strategies using each of the twelve SD weights—0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, and 1.2, in each of the two item pools.

Item pools and item responses

In the 1-PL item pool and the modified 3-PL item pool, there were 171 items, respectively, whose difficulties were equally spaced from -3.4 to 3.4, in interval 0.04. Item discrimination parameters and item pseudo-guessing parameters in each item pool were constants. In the 1-PL item pool, all $a_i = 1$, $c_i = 0$; in the modified 3-PL item pool, all $a_i = 1.25$, $c_i = .25$.

When an item was provided to a simulee whose ability was θ, a random number g generated from a distribution uniform in interval $(0, 1)$ was compared with the item response function $P_i(\theta)$ (see Equation 1.3 and Equation 1.1). Item response $\mu_i = 1$ when g was smaller than or equal to $P_i(\theta)$; otherwise, $\mu_i = 0$. The constant D used in the 1-PL and the modified 3-PL item pools was 1.702 on the IBM compatible microcomputers, was 1.7 on the UNIX workstations.

CAT strategies

In all three CAT strategies—GSSS, DSS, and ZSS—all simulees were assumed a pre-ability estimate $\theta = 0$. The current ability estimate is assumed to be equal to the ability level of a testing point, if the obtained score does not exceed the confidence interval of the expected score at that testing point. The locally best weight (see Equation 1.5) was used in calculating the obtained score and expected score. On the IBM compatible microcomputers, the item selection method was the quasi-match m_i to $\hat{\theta}$ item selection. On the UNIX workstations, the item selection method was the maximum information item selection. The final ability estimation was the MLE of ability. On the IBM compatible microcomputers, Modified Newton-Raphson method (Hamming, 1973, pp. 68-72) was used in the final ability estimation to solve the MLE of ability, after 20 items were administered. On the UNIX workstations, dichotomous search was used to solve MLE of ability. Iteration process continued until the difference between two successive estimated θ values was less than 0.001. In the final ability estimation, if a simulee's item responses were all 1s or all 0s, or a simulee's MLE solution was less than -10, or greater than 10, it would be excluded from the results. Each excluded case was replaced by a simulee with the same ability level, to maintain 100 simulees in each group.

DSS. In DSS, the original search region was [-3.4, 3.4] for all simulees. The sizes of successive search regions was reduced by ratio .5. Each midpoint of successive dichotomous search regions was a testing point. After each item was administered, hypothesis tests were conducted, starting from the testing point of the original search region, until a testing point was found, at which the confidence interval of expected

score $\mu_{x|\theta}$, covered the obtained score. The locally best weight was used to weight the obtained and the expected scores. Whenever a simulee's obtained score was less than a portion of score assumed from random guessing, the obtained score was adjusted using Equation 1.22. Item selection method was either the quasi-match m_i to $\hat{\theta}$ item selection (on the IBM microcomputers) or the maximum information item selection (on the UNIX workstations). In situation where the search region was reduced to less than 0.08, the simulee's current ability was assumed to be equal to the last testing point (the midpoint of the search region), regardless of the hypothesis testing results. The current ability estimate was limited by the original search region [-3.4, 3.4]. The termination criterion for each CAT was fixed number of items. Each CAT terminated after 20 items were administered. The final ability estimate was the MLE of ability, which is not limited by the original search region.

GSSS. GSSS was the same as DSS, except for the ratio of sizes of successive search regions. In GSSS, search region was reduced by ratio $\frac{\sqrt{5}-1}{2} \approx .618$.

ZSS. ZSS (using SD weight) was the same in every respect as DSS or GSSS, except for the determination of testing points. After each item was administered, the obtained score was compared with the expected score at a testing point, which was the same as the previous ability estimate. If the obtained score was within a confidence interval of the expected score, the current ability estimate was the same as the previous estimate of ability. If the obtained score was outside the confidence interval of the expected score, next testing point would be calculated according to Equation 1.26 or Equation 1.28. The hypothesis testing continued until a testing point was found, at which the obtained test score was within the confidence

interval of the expected score. Thus the current ability estimate was assumed to be the same as that testing point. Next item to be chosen was based on the current ability estimate. The locally best weight was used to weight the obtained and the expected scores. Whenever a simulee's obtained score was less than a portion of score assuming from random guessing, the obtained score was adjusted using Equation 1.22. Item selection method was either the quasi-match m_i to $\hat{\theta}$ item selection (on the IBM microcomputers) or the maximum information item selection (on the UNIX workstations).

In ZSS (using SD weight), during the item selection process, if a simulee's current ability estimate was bigger than (or smaller than) the b value of the most difficult item (or the easiest item) (which is 3.4 (or -3.4) in the two item pools), the simulee's current ability estimate was assigned to be 3.4 (or -3.4). Test terminated after 20 items were administered. The final ability estimate was the MLE.

Results

On the IBM compatible microcomputers

Table 2.1 shows the measurement precision for GSSS, DSS, and ZSS using four different weights, in the 1-PL and the modified 3-PL item pools. Value in each cell except for the last column represents the average bias, absolute errors, MSE, and information for 2500 simulees grouped in 25 ability levels. The value in the last column is the replacements made for each CAT strategy using each SD weight, in each of the item pools.

Table 2.1: Measurement precision for GSSS, DSS, and ZSS using four SD weights, in the 1-PL and the modified 3-PL item pools[a]

	CAT strategy	SD weight	Bias	ABE	MSE	Info	Replace
1-PL							
	GSSS						
		0.4	0.012	0.229	0.084	11.951	0
		0.7	0.012	0.226	0.082	11.974	0
		1.0	0.009	0.233	0.088	11.709	0
		1.3	-0.001	0.231	0.085	11.572	0
	DSS						
		0.4	0.011	0.228	0.083	11.945	0
		0.7	0.011	0.229	0.083	11.945	0
		1.0	-0.001	0.229	0.082	11.742	0
		1.3	0.002	0.231	0.080	11.643	0
	ZSS						
		0.4	0.010	0.230	0.085	11.898	0
		0.7	0.014	0.228	0.084	11.856	0
		1.0	0.011	0.235	0.088	11.573	0
		1.3	0.000	0.228	0.082	11.567	0

Table 2.1: (Continued)

CAT strategy	*SD* weight	Bias	ABE	MSE	Info	Replace
Modified 3-PL						
GSSS						
	0.4	0.053	0.283	0.156	9.652	4
	0.7	0.046	0.272	0.136	9.742	5
	1.0	0.039	0.272	0.127	9.469	4
	1.3	0.036	0.274	0.128	9.206	4
DSS						
	0.4	0.051	0.288	0.165	9.535	4
	0.7	0.046	0.269	0.132	9.710	1
	1.0	0.043	0.270	0.129	9.542	3
	1.3	0.040	0.263	0.119	9.390	4
ZSS						
	0.4	0.050	0.270	0.135	9.729	4
	0.7	0.050	0.268	0.130	9.659	4
	1.0	0.043	0.276	0.133	9.215	3
	1.3	0.039	0.280	0.135	8.953	5

[a]2500 simulees in each cell.

Bias, absolute errors, and frequencies of replacement. From Table 2.1, one can see that no matter what SD weight was used, the average bias for GSSS, DSS, and ZSS were very small in the 1-PL model. The average bias for the three CAT strategies were also small in the modified 3-PL model, but were slightly bigger than those in the 1-PL model. There were not many differences among the mean absolute errors obtained by using different SD weights in the 1-PL item pool for all the three CAT strategies. In the modified 3-PL item pool, Smaller mean absolute errors were obtained when SD weights were 0.7 and 1.0 for GSSS; were 0.7, 1.0, and 1.3 for DSS; were 0.4 and 0.7 for ZSS. No replacement was made in the 1-PL item pool for the three CAT strategies. In the modified 3-PL item pool, a few simulees had negative infinity final MLE ability estimates. Those simulees were replaced by simulees with the same ability levels. The total frequencies of replacement in each CAT strategy using each SD weight were small.

MSE and information for GSSS. Table 2.1 lists the MSEs and test information for each CAT in each item pool. In the 1-PL item pool, the MSEs for GSSS using SD weights 0.7, 0.4, 1.3, and 1.0, were: 0.082, 0.084, 0.085, and 0.088, respectively. From Figure 2.1, one can see that the MSEs for GSSS using four SD weights are not different. The means of test information for GSSS using SD weights 0.7, 0.4, 1.0, and 1.3, were: 11.974, 11.951, 11.709, and 11.572, respectively. Figure 2.2 shows the information curves for GSSS using four different SD weights. In the middle range of ability, information obtained by using different SD weights was almost the same. In the lower ability or higher ability levels, information obtained using SD weights 0.7 and 0.4 was much higher than those obtained using SD weights 1.0 and 1.3.

In the modified 3-PL item pool, the MSEs for GSSS using SD weights 1.0, 1.3, 0.7, and 0.4, were: 0.127, 0.128, 0.136, and 0.156, respectively. From Figure 2.3, one can see that the MSEs for GSSS using four SD weights are different. The biggest MSE was found by using SD weight 0.4. The means of test information for GSSS using SD weights 0.7, 0.4, 1.0, and 1.3, were: 9.742, 9.652, 9.469, and 9.206, respectively. Figure 2.4 shows the information curves for GSSS using four different SD weights. In the middle range of ability, information obtained by using different SD weights was almost the same. In the lower ability or higher ability levels, information obtained using SD weights 0.7 and 0.4 was much higher than those obtained using SD weights 1.0 and 1.3.

The best SD weights for GSSS seemed to be 0.7 in the 1-PL item pool, 0.7 and 1.0 in the modified 3-PL item pool. Generally speaking, using SD weight 0.7 for GSSS could provide optimal measurement precision in both item pools.

MSE and information for DSS. In the 1-PL item pool, the MSEs for DSS using SD weights 1.3, 1.0, 0.7, and 0.4, were: 0.080, 0.082, 0.083, and 0.083, respectively. From Figure 2.5, one can see that the MSEs for DSS using four SD weights are not different. The means of test information for DSS using SD weights 0.4, 0.7, 1.0, and 1.3, were: 11.945, 11.945, 11.742, and 11.643, respectively. Figure 2.6 shows the information curves for DSS using different SD weights. In the middle range of ability, information obtained by using different SD weights was almost the same. In the extremely lower ability or extremely higher ability levels, information obtained using SD weights 0.7 and 0.4 was higher than those obtained using SD weights 1.0 and 1.3. The information curves were more flat compared to those obtained by GSSS,

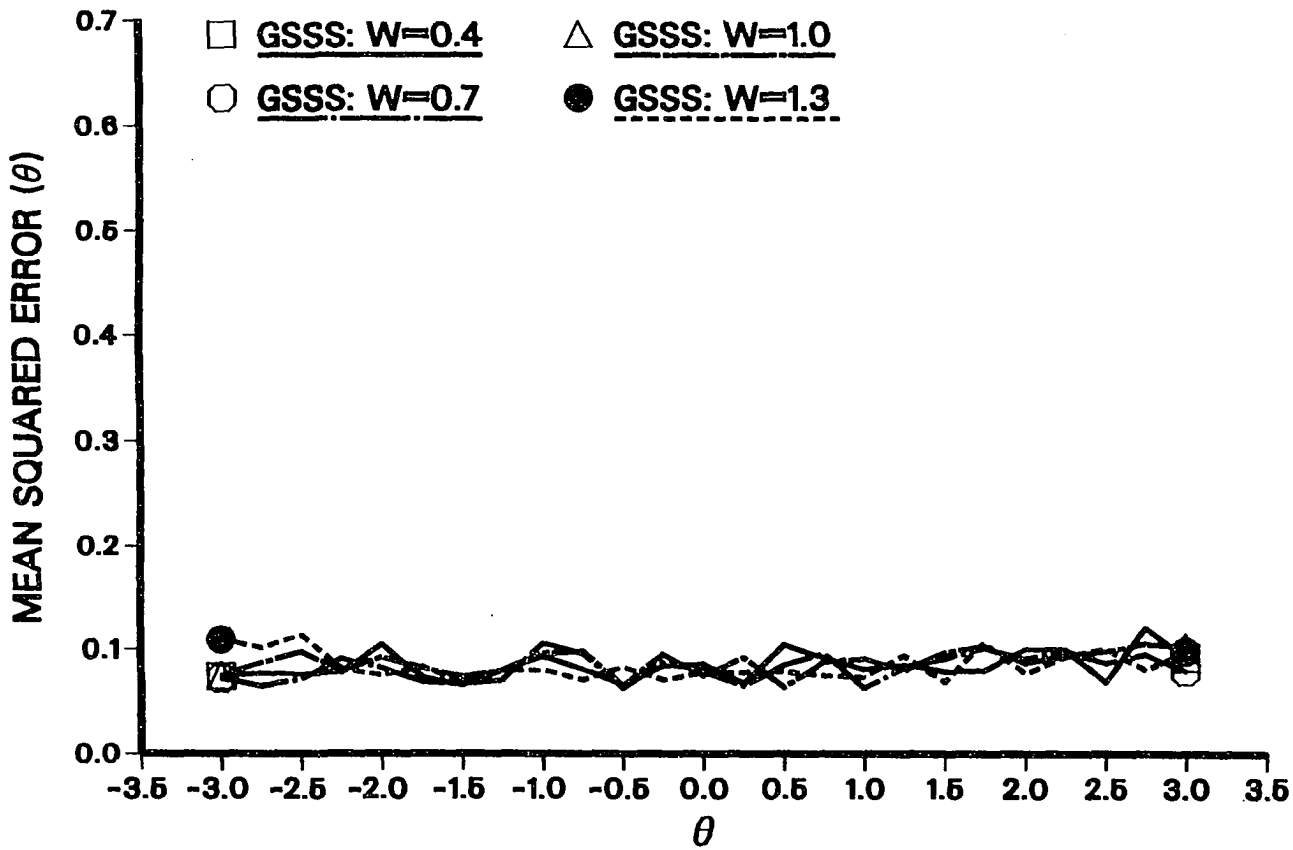

Figure 2.1: Mean squared errors for GSSS using four SD weights (W) in the 1-PL item pool

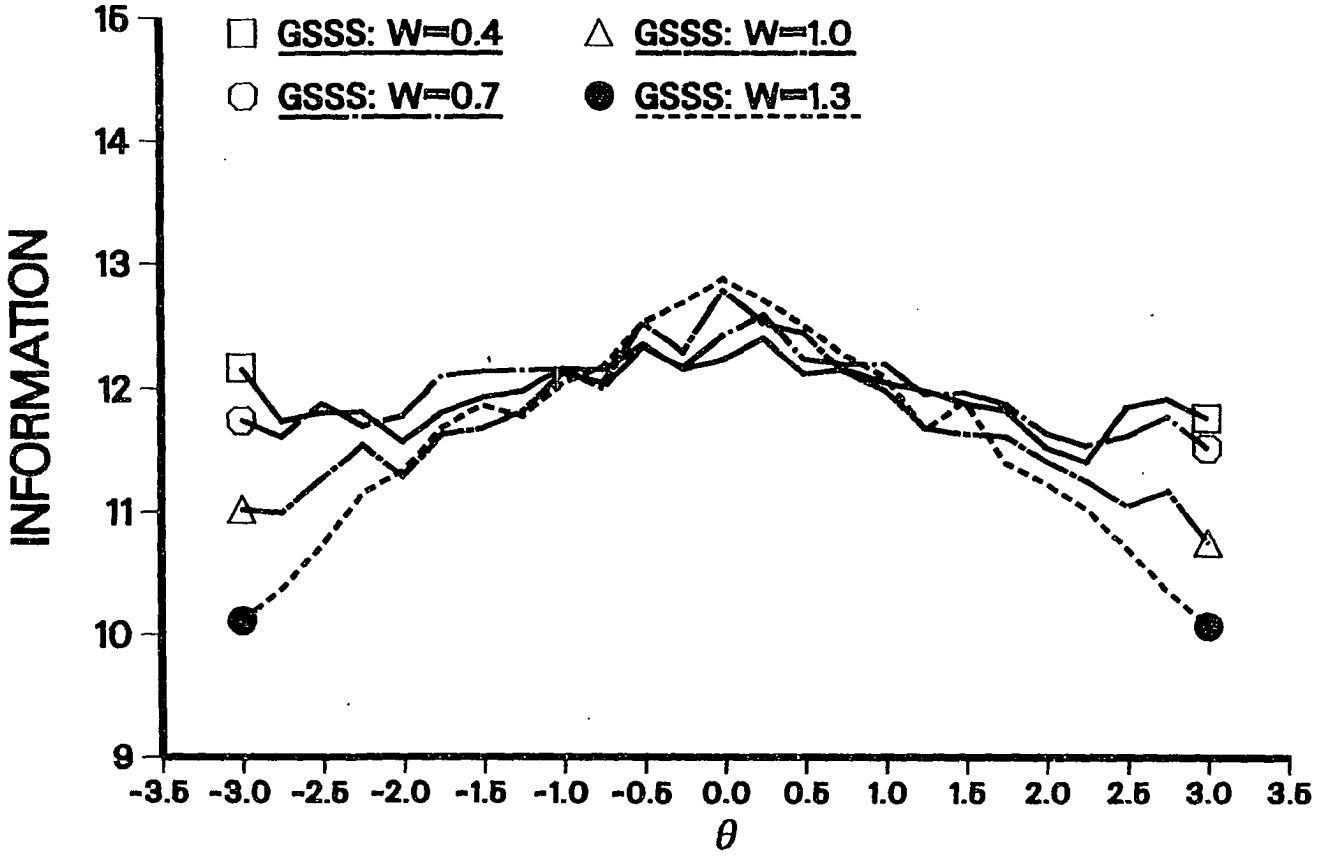

Figure 2.2: Test information for GSSS using four SD weights (W) in the 1-PL item pool

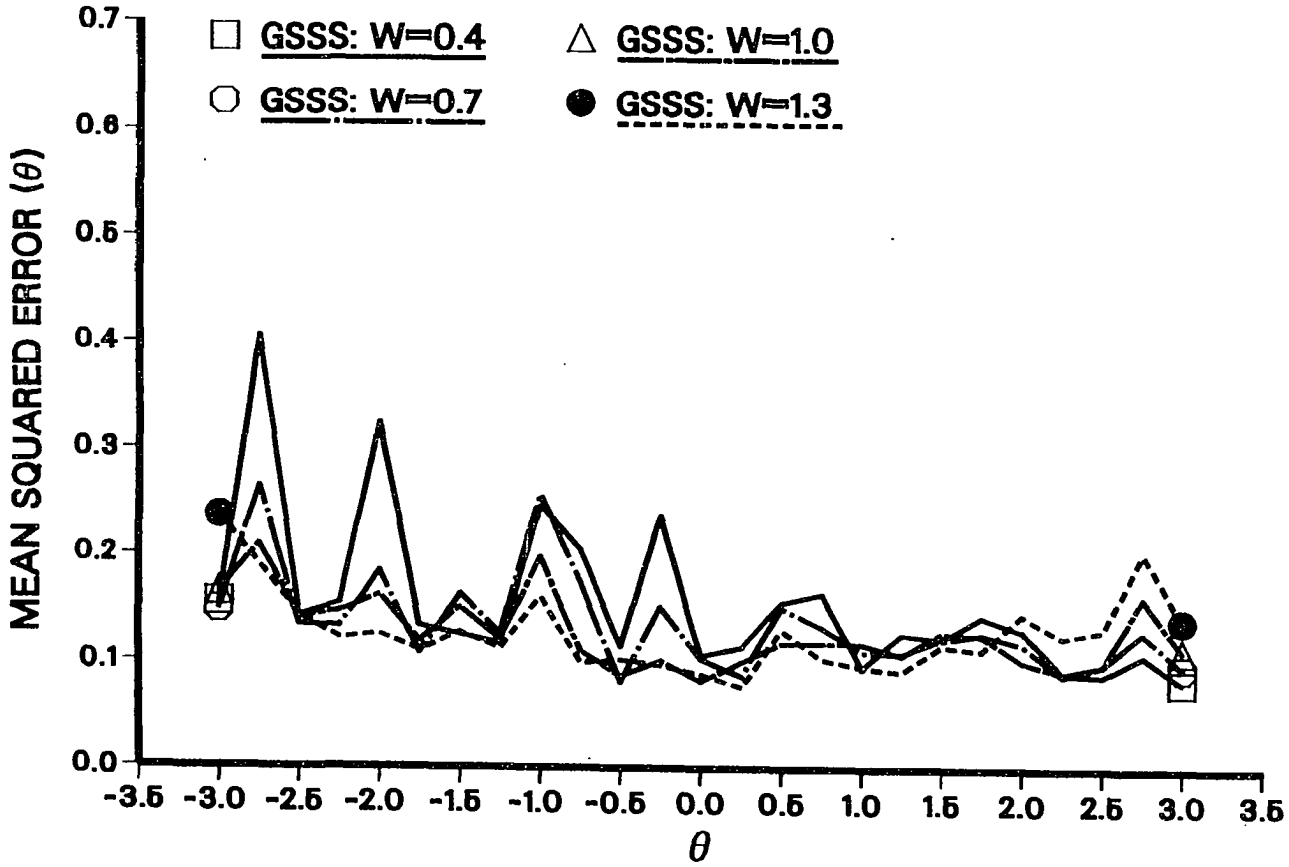

Figure 2.3: Mean squared errors for GSSS using four SD weights (W) in the modified 3-PL item pool

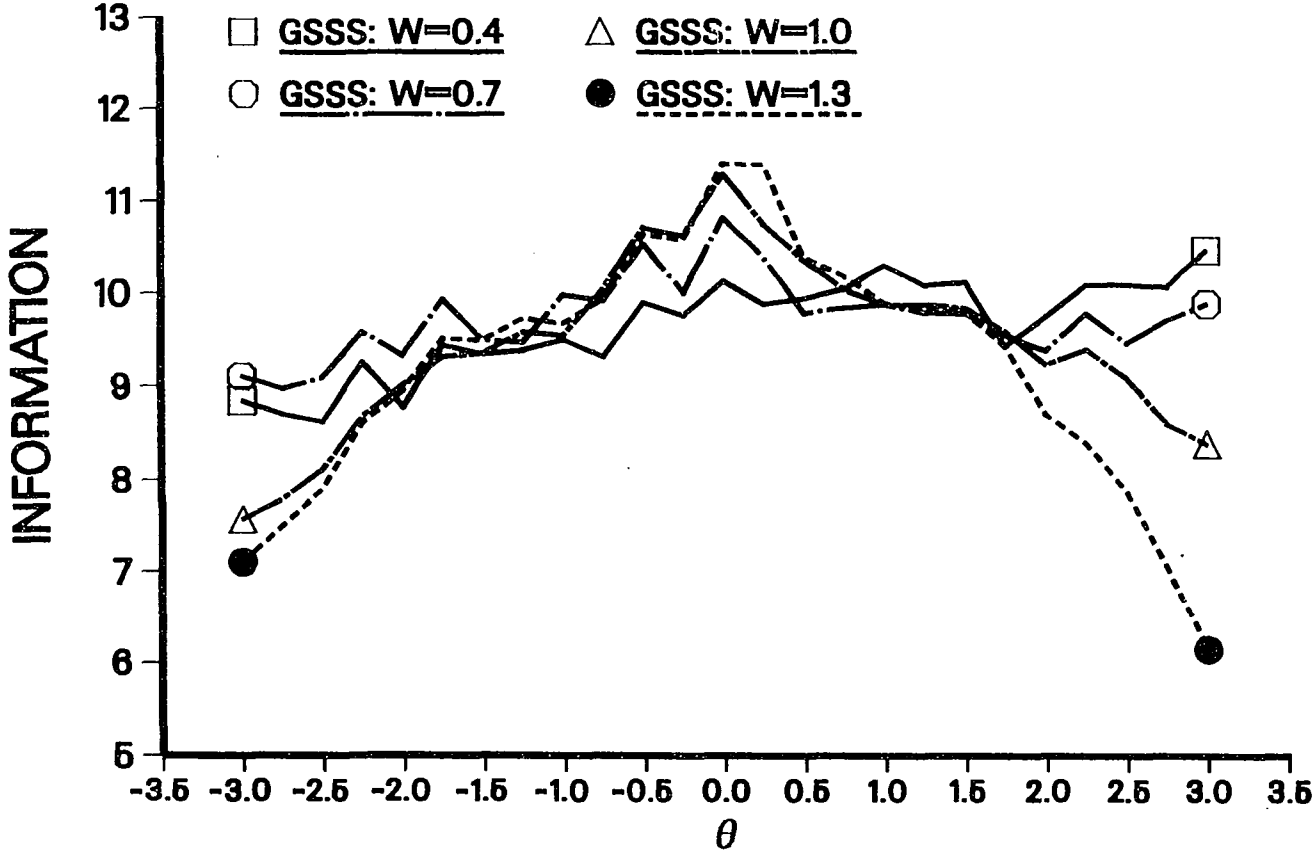

Figure 2.4: Test information for GSSS using four SD weights (W) in the modified 3-PL item pool

which were slightly more bell shaped.

In the modified 3-PL item pool, the MSEs for DSS using SD weights 1.3, 1.0, 0.7, and 0.4, were: 0.119, 0.129, 0.132, and 0.165, respectively. Figure 2.7 shows that the MSEs for DSS using four SD weights are different. The greatest MSE was generated by using SD weight 0.4. The means of test information for DSS using SD weights 0.7, 1.0, 0.4, and 1.3, were: 9.710, 9.542, 9.535, and 9.390, respectively. Figure 2.8 shows the information curves for DSS using four different SD weights. In the lower and middle range of ability, information obtained by using different SD weights was almost the same, except for using SD weight 0.4: that provided lower information. In the extremely higher ability levels, information obtained using SD weights 0.4 and 0.7 was higher than that obtained using SD weights 1.0 and 1.3. The information curves were slightly more flat compared to those obtained by GSSS.

The best SD weights for DSS seemed to be 0.4, 0.7, and 1.0 in the 1-PL item pool, and 0.7 and 1.0 for the modified 3-PL item pool. In short, using SD weight 0.7 and 1.0 for DSS could provide optimal measurement precision in both item pools.

MSE and information for ZSS. In the 1-PL item pool, the MSEs for ZSS using SD weights 1.3, 0.7, 0.4, and 1.0, were: 0.082, 0.084, 0.085, and 0.088, respectively. From Figure 2.9, one can see that the MSEs for ZSS using four SD weights are not different. The means of test information for ZSS using SD weights 0.4, 0.7, 1.0, and 1.3, were: 11.898, 11.856, 11.573, and 11.567, respectively. Figure 2.10 shows the information curves for ZSS using four different SD weights. In the middle range of ability, information obtained by using different SD weights was almost the same. In the lower ability or higher ability levels, information obtained

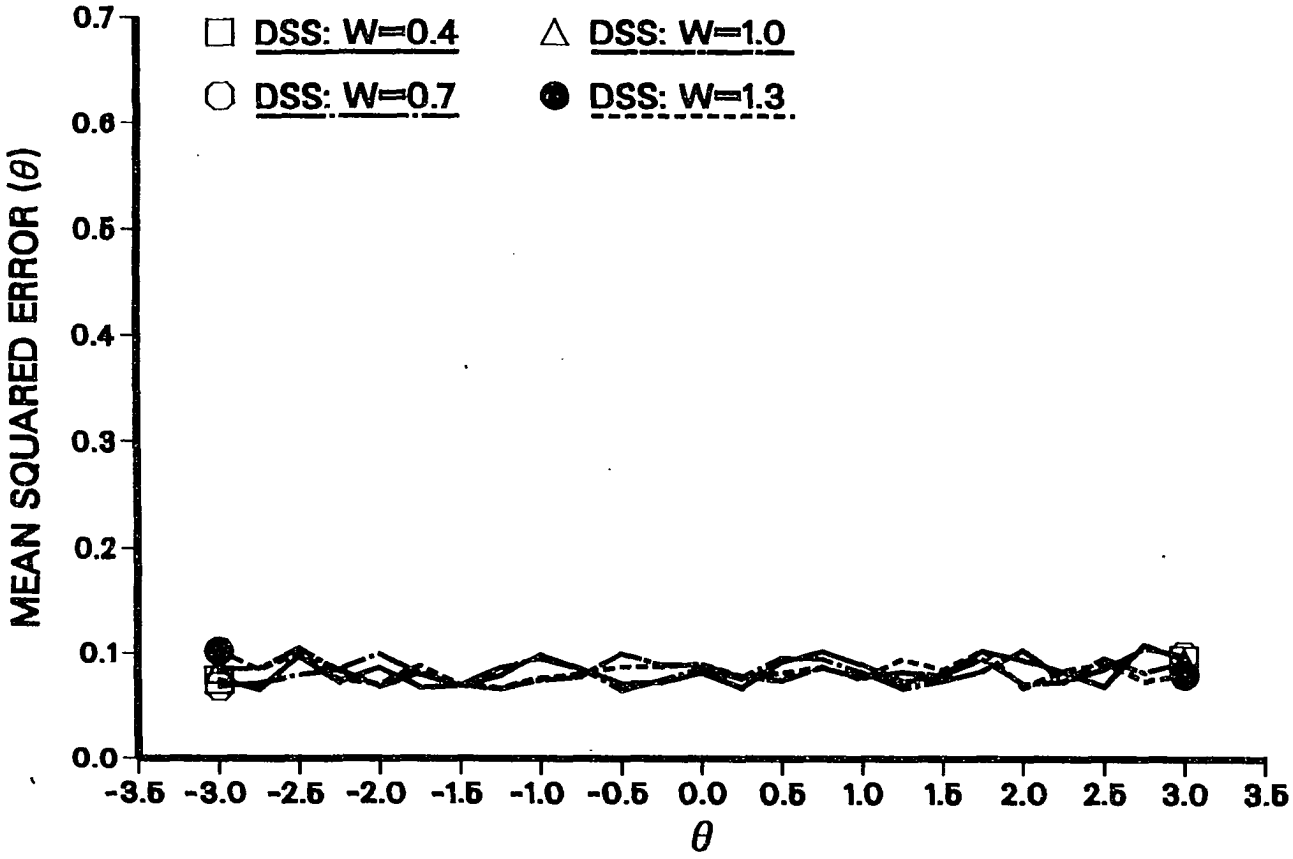

Figure 2.5: Mean squared errors for DSS using four SD weights (W) in the 1-PL item pool

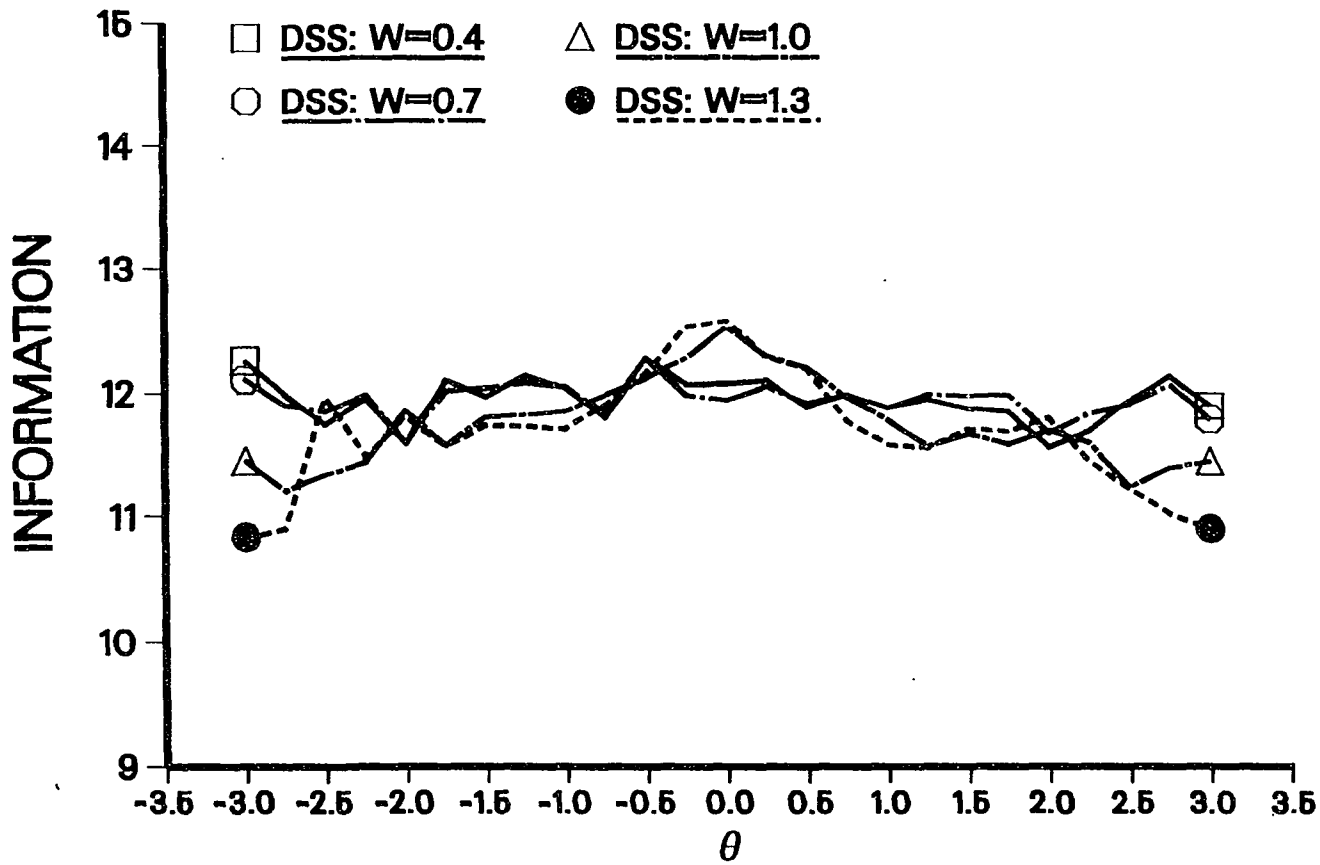

Figure 2.6: Test information for DSS using four SD weights (W) in the 1-PL item pool

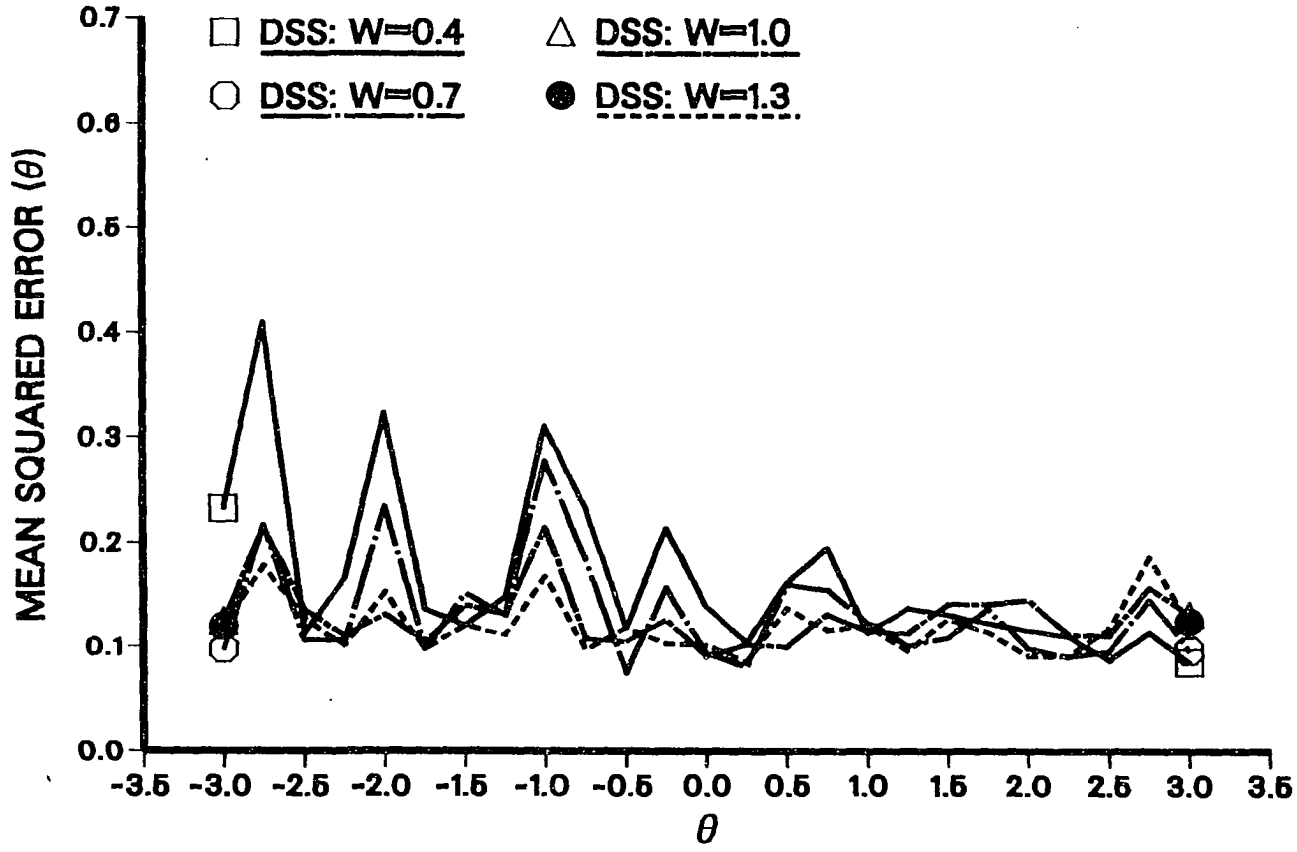

Figure 2.7: Mean squared errors for DSS using four SD weights (W) in the modified 3-PL item pool

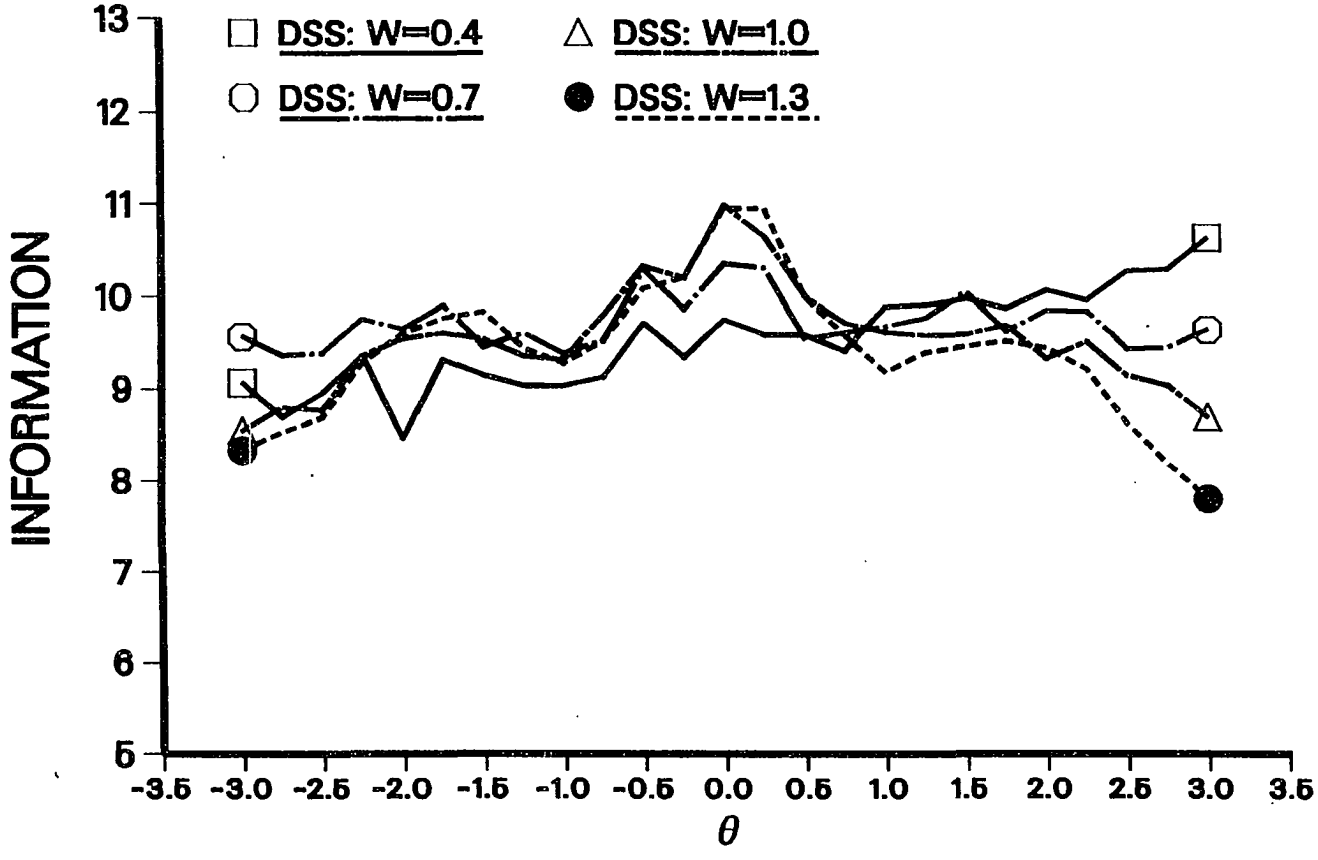

Figure 2.8: Test information for DSS using four SD weights (W) in the modified 3-PL item pool

using SD weights 0.7 and 0.4 was much higher than those obtained using SD weights 1.0 and 1.3.

In the modified 3-PL item pool, the MSEs for ZSS using SD weights 0.7, 1.0, 1.3, and 0.4, were: 0.130, 0.133, 0.135, and 0.135, respectively. Figure 2.11 shows that MSEs for ZSS using four SD weights are not different. The means of test information for ZSS using SD weights 0.4, 0.7, 1.0, and 1.3, were: 9.729, 9.659, 9.215, and 8.953, respectively. Figure 2.12 shows the information curves for ZSS using four different SD weights. In a small range of middle ability levels, information obtained by using different SD weights was almost the same. In ability levels other than the middle range, information obtained by using using SD weights 0.7 and 0.4 was much higher than that obtained by using SD weights 1.0 and 1.3.

The best SD weights for ZSS seemed to be 0.4 and 0.7 in both the 1-PL and the 3-PL item pools.

On the Digital UNIX workstations

Tables 2.2, 2.3, and 2.4 show the measurement precision for GSSS, DSS, and ZSS using twelve different weights in the 1-PL and the modified 3-PL item pools. Value in each cell except for the last column represents the average bias, absolute errors, MSE, and information for 2500 simulees grouped in 25 ability levels. The value in the last column is the replacements made for each CAT strategy using each SD weight, in each of the item pools.

Bias and frequencies of replacement. From Tables 2.2, 2.3, and 2.4, one can see that no matter what SD weight was used, the average bias for GSSS, DSS, and

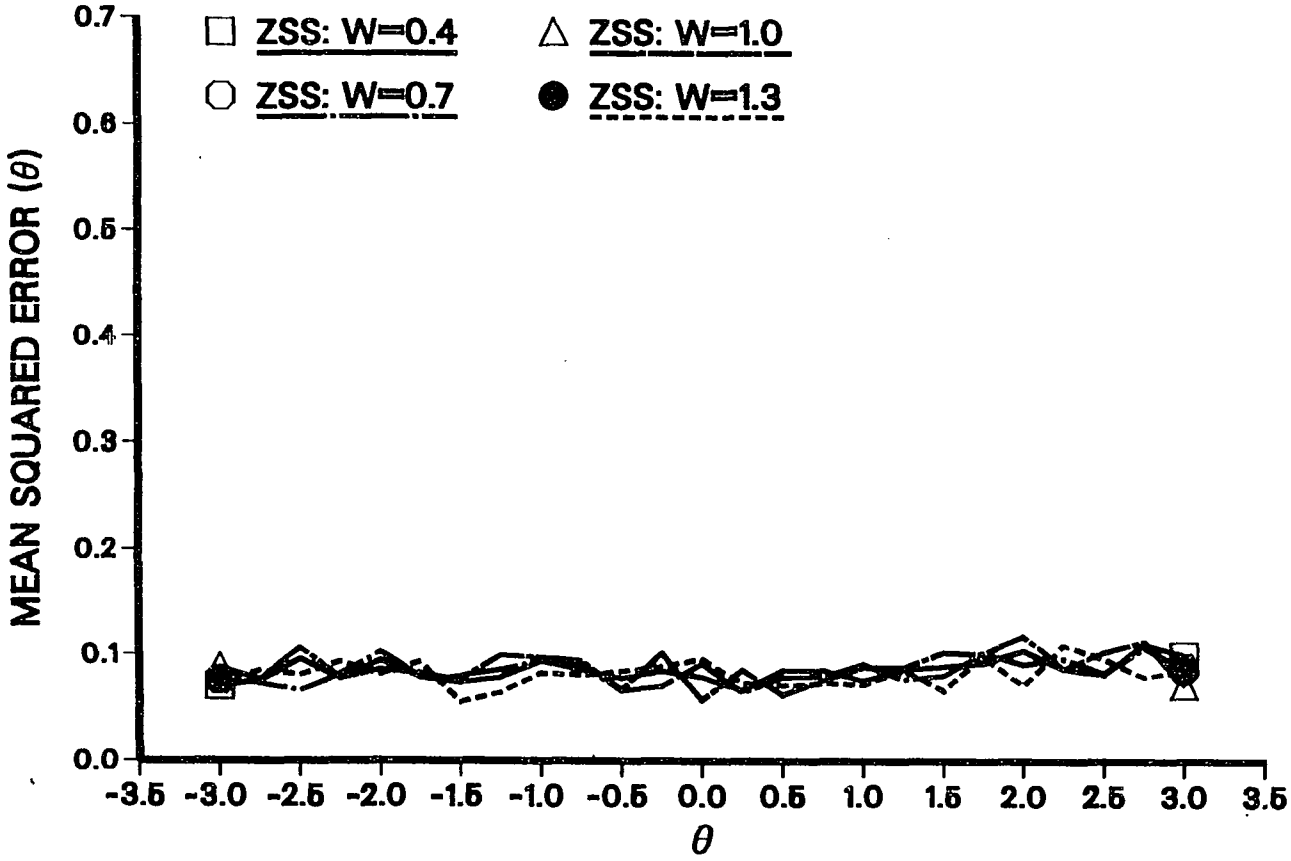

Figure 2.9: Mean squared errors for ZSS using four SD weights (W) in the 1-PL item pool

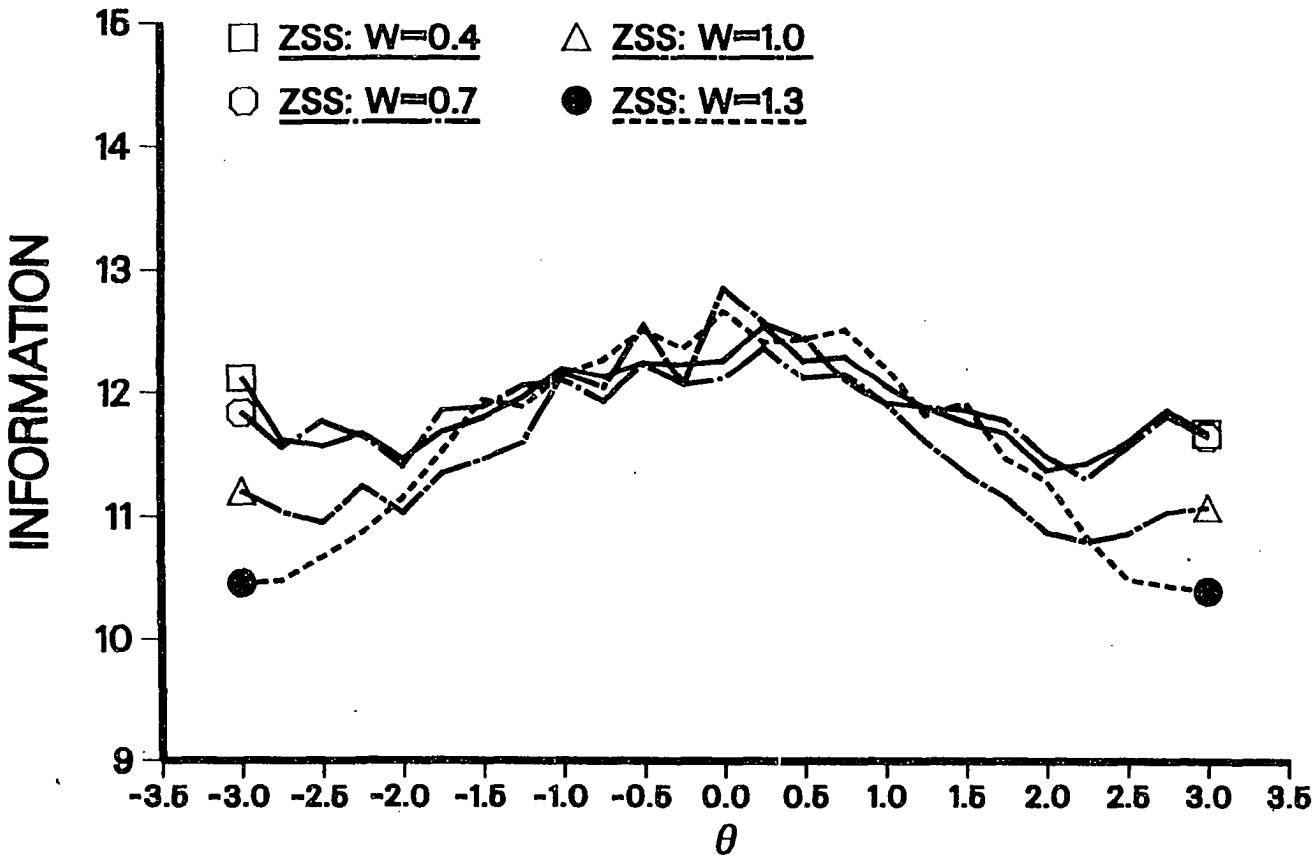

Figure 2.10: Test information for ZSS using four SD weights (W) in the 1-PL item pool

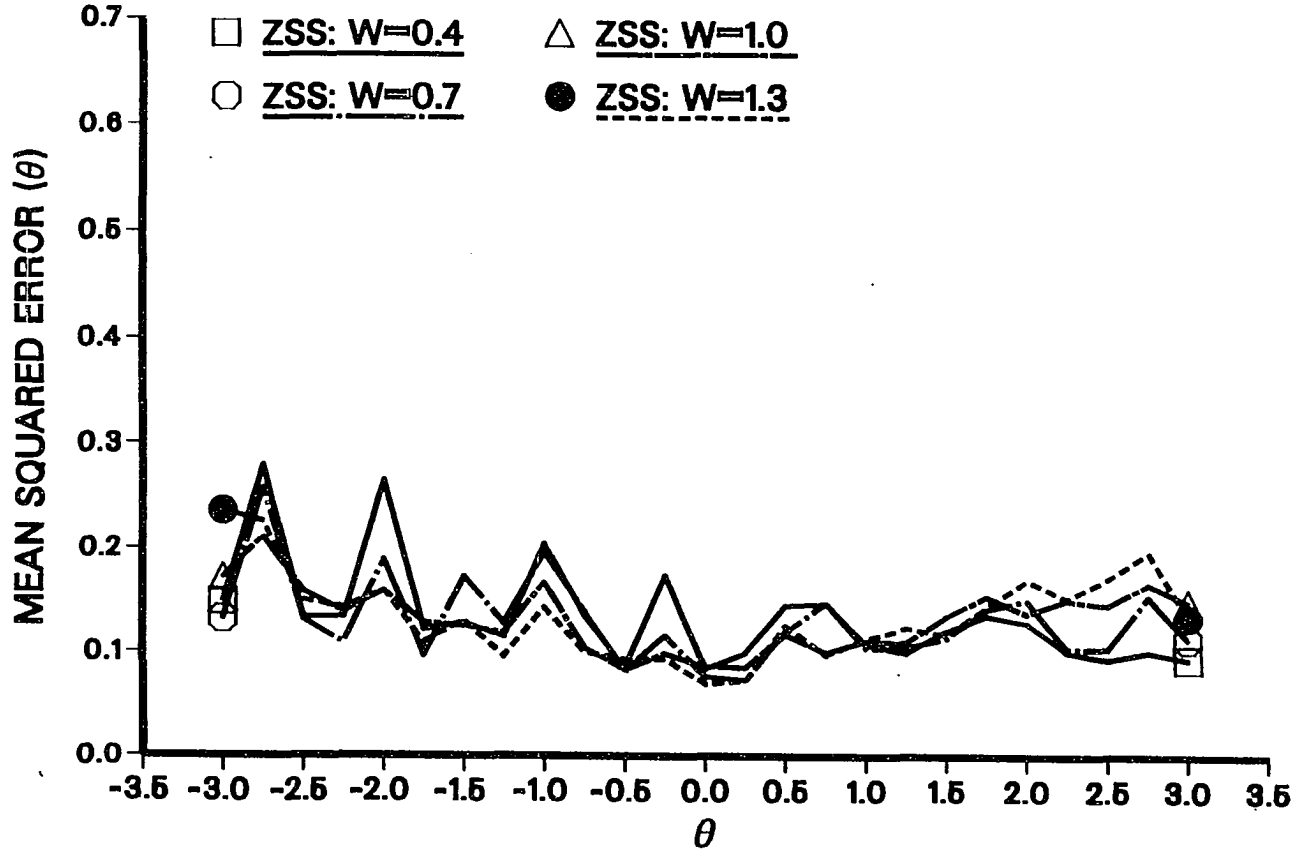

Figure 2.11: Mean squared errors for ZSS using four SD weights (W) in the modified 3-PL item pool

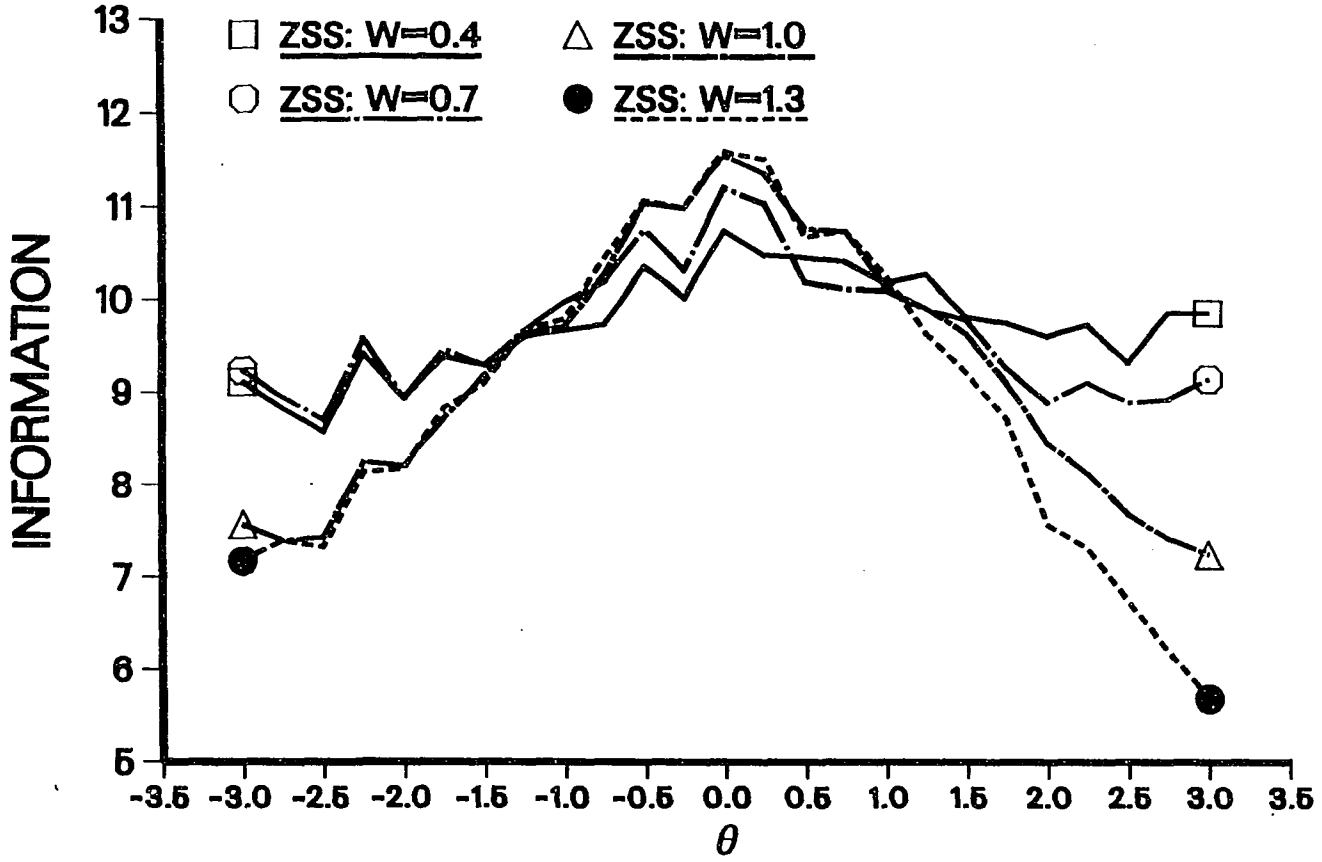

Figure 2.12: Test information for ZSS using four SD weights (W) in the modified 3-PL item pool

ZSS was very small in the 1-PL model. The average bias for the three CAT strategies was also small in the modified 3-PL model, but was slightly bigger than those in the 1-PL model. No replacement was made for all three CAT strategies using any of the SD weights in the 1-PL item pool. In the modified 3-PL item pool, a few simulees had negative infinity final MLE ability estimates. Those simulees were replaced by simulees with the same ability levels. The total frequencies of replacement in each CAT strategy, using each SD weight, were small, especially when SD weights were not bigger than 0.9.

Bias, absolute errors, MSE, and information for GSSS. Table 2.2 also shows the average absolute errors, MSE, and information for GSSS using twelve SD weights in each item pool. From Table 2.2 one can see that there was a wide range of SD weights that could provide higher measurement precision for GSSS. In the 1-PL item pool, using SD weights from 0.4 to 0.9 measured very well. Using SD weights smaller than 0.4 or bigger than 0.9 could produce less accurate and less efficient measurement result than did using SD weights in the middle range. Using SD weights 0.7 and 0.8 could provide the most precise measurement results. In the modified 3-PL item pool, using SD weights from 0.5 to 0.9 measured very well. Using SD weights smaller than 0.5 or bigger than 0.9 resulted in less precise ability estimate than did using the SD weights in the middle range. Using SD weights 0.6, 0.7, and 0.8 could produce the most precise ability estimates. In short, using SD weights 0.7 and 0.8 for GSSS could provide optimal measurement precision in both item pools.

Bias, absolute errors, MSE, and information for DSS. Table 2.3 also shows the average absolute errors, MSE, and information for DSS using twelve SD

Table 2.2: Measurement precision for GSSS using twelve SD weights, in the 1-PL and the modified 3-PL item pools[a]

Item pool	SD weight	Bias	ABE	MSE	Info	Replace
1-PL						
	0.1	-0.018	0.232	0.087	11.663	0
	0.2	-0.017	0.234	0.088	11.809	0
	0.3	-0.013	0.234	0.088	11.811	0
	0.4	-0.009	0.229	0.085	11.914	0
	0.5	-0.007	0.227	0.084	11.896	0
	0.6	-0.013	0.232	0.087	11.877	0
	0.7	-0.011	0.228	0.084	11.831	0
	0.8	-0.006	0.228	0.084	11.816	0
	0.9	-0.006	0.232	0.086	11.769	0
	1.0	-0.004	0.235	0.088	11.597	0
	1.1	-0.007	0.238	0.091	11.358	0
	1.2	-0.005	0.239	0.091	11.228	0
Modified 3-PL						
	0.1	0.057	0.294	0.184	9.321	2
	0.2	0.043	0.282	0.164	9.531	3
	0.3	0.043	0.282	0.162	9.518	2
	0.4	0.045	0.276	0.152	9.582	5
	0.5	0.040	0.272	0.145	9.642	3
	0.6	0.035	0.267	0.138	9.629	2
	0.7	0.028	0.268	0.132	9.581	2
	0.8	0.029	0.270	0.128	9.514	2
	0.9	0.026	0.268	0.128	9.430	3
	1.0	0.029	0.271	0.123	9.009	5
	1.1	0.027	0.275	0.128	8.851	6
	1.2	0.028	0.282	0.136	8.586	7

[a] 2500 simulees in each cell.

Table 2.3: Measurement precision for DSS using twelve SD weights, in the 1-PL and the modified 3-PL item pools[a]

Item pool	SD weight	Bias	ABE	MSE	Info	Replace
1-PL						
	0.1	-0.017	0.232	0.087	11.655	0
	0.2	-0.014	0.228	0.085	11.750	0
	0.3	-0.008	0.232	0.087	11.863	0
	0.4	-0.014	0.231	0.085	11.868	0
	0.5	-0.012	0.231	0.086	11.842	0
	0.6	-0.008	0.231	0.086	11.868	0
	0.7	-0.012	0.233	0.086	11.838	0
	0.8	-0.008	0.236	0.088	11.788	0
	0.9	-0.009	0.232	0.086	11.694	0
	1.0	-0.007	0.237	0.090	11.581	0
	1.1	-0.006	0.237	0.089	11.408	0
	1.2	-0.008	0.239	0.090	11.304	0
Modified 3-PL						
	0.1	0.057	0.297	0.192	9.215	3
	0.2	0.048	0.284	0.172	9.434	4
	0.3	0.046	0.282	0.166	9.436	2
	0.4	0.042	0.279	0.158	9.525	2
	0.5	0.046	0.280	0.157	9.509	4
	0.6	0.037	0.277	0.150	9.510	1
	0.7	0.029	0.273	0.133	9.549	1
	0.8	0.030	0.271	0.131	9.555	3
	0.9	0.029	0.270	0.132	9.435	0
	1.0	0.028	0.274	0.125	9.142	1
	1.1	0.026	0.277	0.129	9.033	0
	1.2	0.031	0.273	0.128	8.830	5

[a] 2500 simulees in each cell.

Table 2.4: Measurement precision for ZSS with different SD weights, using the 1-PL and the modified 3-PL item pools[a]

Item pool	SD weight	Bias	ABE	MSE	Info	Replace
1-PL						
	0.1	-0.013	0.231	0.086	11.636	0
	0.2	-0.012	0.232	0.086	11.770	0
	0.3	-0.009	0.227	0.085	11.846	0
	0.4	-0.009	0.228	0.084	11.859	0
	0.5	-0.006	0.230	0.086	11.871	0
	0.6	-0.007	0.231	0.086	11.799	0
	0.7	-0.008	0.230	0.086	11.733	0
	0.8	-0.009	0.234	0.089	11.695	0
	0.9	-0.009	0.233	0.088	11.656	0
	1.0	-0.012	0.236	0.091	11.450	0
	1.1	-0.007	0.242	0.094	11.196	0
	1.2	-0.004	0.240	0.093	11.197	0
Modified 3-PL						
	0.1	0.058	0.293	0.190	9.237	3
	0.2	0.045	0.286	0.172	9.465	4
	0.3	0.048	0.284	0.168	9.474	4
	0.4	0.035	0.271	0.137	9.700	5
	0.5	0.035	0.275	0.141	9.672	5
	0.6	0.035	0.273	0.142	9.626	6
	0.7	0.028	0.271	0.137	9.506	3
	0.8	0.024	0.268	0.129	9.484	3
	0.9	0.026	0.265	0.125	9.401	4
	1.0	0.032	0.281	0.138	8.862	9
	1.1	0.036	0.282	0.141	8.644	10
	1.2	0.034	0.290	0.147	8.442	14

[a]2500 simulees in each cell.

weights in each item pool. From Table 2.3 one can see that there was a wide range of SD weights that could provide higher measurement precision for DSS. In the 1-PL item pool, using SD weights from 0.1 to 0.9 measured very well. Using SD weights bigger than 0.9 measured less accurately and less efficiently than did using SD weights in the lower range. Using SD weights 0.4, 0.5, 0.6 and 0.7 provided the most precise ability estimates. In the modified 3-PL item pool, using SD weights from 0.7 to to 1.0 measured very well. Using SD weights smaller than 0.7 or bigger than 1.0 resulted in less precise ability estimate than did using SD weights in the middle range. Using SD weight 0.7 and 0.8 provided the most precise ability estimates. In short, using SD weight 0.7 for DSS could provide optimal measurement precision in both item pools.

Bias, absolute errors, MSE, and information for ZSS. Table 2.4 also shows the average absolute errors, MSE, and information for ZSS using twelve SD weights in each item pool. From Table 2.4 one can see that there was a wide range of SD weights that could provide higher measurement precision for ZSS. In the 1-PL item pool, using SD weights from 0.1 to 0.7 measured very well. Using SD weights bigger than 0.7 measured less accurately and less efficiently than did using SD weights in the lower range. Using SD weights 0.4, 0.5, 0.6 and 0.7 could provide the most precise ability estimates. In the modified 3-PL item pool, using SD weights from 0.4 to to 0.9 measured very well. Using SD weights smaller than 0.4 or bigger than 0.9 resulted in less precise ability estimate than did using SD weights in the middle range. Using SD weight 0.4, 0.5, 0.6, 0.7 and 0.8 could provide the most precise ability estimates. Using SD weight 0.4, 0.5, 0.6, and 0.7 for ZSS could provide

optimal measurement precision in both item pools.

Discussion

Results of Study One showed that the best SD weight was around 0.7 for GSSS and DSS, were from 0.4 to 0.7 for ZSS in both of the 1-PL and the modified 3-PL item pools on the microcomputers and the UNIX workstations. The results of the present study suggested that a range of SD weights could be used in GSSS, DSS, and ZSS, in various item pools, to achieve accurate measurement results.

In the 1-PL model, the average measurement bias was very small. In the modified 3-PL model, bias was generally greater than those in the 1-PL model, but still in a reasonably small range. The absolute errors and MSE are confounded with the deviation from the mean of the ability estimates and the deviation of that mean from the true latent ability level. Since the MLE of ability was used in the final ability estimate in each strategy, it was expected that the MLE of ability was unbiased and the bias did not take an important role in affecting the precision of measurement.

To understand why SD weight around 0.7 can provide a more precise ability estimate for the three CAT strategies, one should recall the nature of the current ability estimation in GSSS, DSS, and ZSS. In those CAT strategies, after each item is selected based on the current ability estimate, the obtained score x at each successive testing point is calculated. The expected score of x ($\mu_{x|\theta}$), and the variance of x ($\sigma^2_{x|\theta}$), at each testing point are also calculated. A confidence interval of the expected score is defined by ($\mu_{x|\theta} \pm Z_\alpha \sigma_{x|\theta}$), where Z_α is called SD weight in the present research, which defined the size of the confidence interval. As mentioned previously, if the locally best weight (see Equation 1.5) is used in calculating x, test

score can provide an ability estimate with optimal or nearly optimal precision. With increasing items, the locally best weight score tends to be normally distributed, with mean $\mu_{x|\theta}$ and variance $\sigma^2_{x|\theta}$ (Birnbaum, 1968). Therefore, if the number of items administered is large, according to the normal distribution, one can approximate that the confidence intervals of the current ability estimate defined by SD weights 0.4, 0.7, 1.0, and 1.3 (Z_α = 0.4, 0.7, 1.0, and 1.3), in order, will approximately have 31%, 52%, 64%, and 83% chances to cover the test score x. In other words, the test score will have 31%, 52%, 64%, or 83% chance to fall into the confidence interval, if the SD weights of 0.4, 0.7, 1.0, or 1.3 are used, respectively. The chance of a type I error is 69%, 48%, 36%, and 17%, respectively. The chance of a type II error is unknown. But by using SD weights less than 0.7 (or bigger than 0.7), the chance of a type II error is smaller than (or bigger than) that by using SD weight 0.7.

During the earlier stage of a CAT, the number of items administered is small, the test score distribution is unknown. Norden (1973) indicated that "as there is no complete theory of estimation for small samples, no simple conclusions capable of a wide application about MLE's, or any other estimator, can be made" (p. 47). By applying a confidence interval in determining the current ability estimate, a more conservative decision will be made. That is, the current ability estimate tends to remain in the previous testing point, unless the test score exceeds the confidence interval of the expected score. Study One showed that SD weight 0.7 provided the most precise ability estimates in both of the 1-PL and the modified 3-PL models. SD weight from 0.4 to 0.7 also performed well for ZSS. A moderate SD weight, say, 0.7 was appropriate for all the three strategies, no matter what item pool was used. If item pool changes, the best set of SD weights may change too. Study One indicates

that a wide range of SD weights can be used in GSSS, DSS, and ZSS CAT, say, from 0.4 to 1.0, depending upon the purpose of the test, the nature of the item pool, and the CAT strategy applied.

Two kinds of computers were used in the present research. Though the executing speed on the UNIX workstations was much faster than that on the microcomputers, and the random number generator on the UNIX workstations functioned slightly better than that on the microcomputers, the present research included results from both kinds of computers since CAT is usually operating on the microcomputers.

CHAPTER 3. STUDY TWO: COMPARISON OF TWO ITEM SELECTION METHODS FOR GSSS, DSS, AND ZSS, AND TWO VERSIONS OF ZSS IN THE 1-PL AND THE 2-PL ITEM POOLS

In Chapter 1, several item selection methods of CAT were described. The quasi-match m_i to $\hat{\theta}$ item selection method is one of the simplest item selection methods. The maximum information item selection method is one of the most sophisticated item selection methods that can select most informative items to be presented. In the 1-PL and the modified 3-PL item pools, the as and cs are constant for all items in each item pool. In item pools like these, the item selected by the quasi-match m_i to $\hat{\theta}$ item selection is the same as that selected by the match m_i to $\hat{\theta}$ item selection and the maximum information item selection, when the current ability estimate is a point estimate; is the same as or slightly easier (or more difficult) than that selected by the match m_i to $\hat{\theta}$ item selection and the maximum information item selection, when the current ability estimate is an interval estimate. For GSSS, DSS, and ZSS (using SD weight), the measurement precision for using the quasi-match m_i to $\hat{\theta}$ item selection and the maximum information item selection should not significantly differ in the 1-PL and the modified 3-PL item pools. The quasi-match m_i to $\hat{\theta}$ item selection method should be more computationally effective than the maximum information item selection method.

Two versions of ZSS are proposed in the present research—ZSS (using SD weight) and ZSS (no SD weight). Statistical hypothesis testing is involved in the former to determine the current ability estimate. A SD weight is used for computing the confidence interval of the expected score at a testing point. However, no statistical hypothesis testing is involved in ZSS (no SD weight) in determining the current ability estimate. The Z-score estimate evaluated at the previous current ability estimate is the current ability estimate. The next item to be presented is based on the current ability estimate. The current ability estimate is an interval estimate in ZSS (using SD weight), is a point estimate in ZSS (no SD weight). Generally speaking, the ability estimates obtained by ZSS using an appropriate SD weight are only slightly different from those obtained by ZSS (no SD weight). In the 1-PL and the 3-PL item pools, ZSS (using SD weight) using the quasi-match m_i to $\hat{\theta}$ item selection should provide as precise ability estimates as those provided by ZSS (no SD weight) using the maximum information item selection.

Design

Monte Carlo studies were conducted to compare the measurement precision and computational efficiency of the quasi-match m_i to $\hat{\theta}$ and the maximum information item selection methods for GSSS, DSS, and ZSS, to compare the measurement precision and computational efficiency of ZSS (using SD weight) and ZSS (no SD weight), in the 1-PL and the modified 3-PL item pools. Each CAT contained 20 items. Four measurement precision indices—bias, absolute errors, MSE, and test information—were calculated. Computer simulated examinees were used. The programs were written in C language (compiled) by the author and running on the

Digital UNIX workstations. The executing time of each program was also recorded for the comparison of the computational efficiency of different item selection methods and CAT strategies.

Method

Simulees

A simulee was a computer generated examinee with a true ability value θ. For the study of comparison of the quasi-match m_i to $\hat{\theta}$ and the maximum information item selection methods, there would be 2500 simulees for each of the three CAT strategies—GSSS, DSS, and ZSS using SD weight 0.7, in the 1-PL and the modified 3-PL item pools, respectively, with 100 as a group at each of 25 ability levels equally spaced in [-3, 3], in interval of 0.25. For the study of comparison of the two versions of ZSS, there would be the same 2500 simulees for ZSS using SD weight 0.4 using the quasi-match m_i to $\hat{\theta}$ item selection and the maximum information item selection, and for ZSS (no SD weight) using the maximum information item selection, in the 1-PL and the modified 3-PL item pools.

Item pools and item selection methods

The 1-PL item pool and the modified 3-PL item pool were the same as the item pools in Study One. Two item selection methods were used in Study Two—the quasi-match m_i to $\hat{\theta}$ item selection and the maximum information item selection.

In the quasi-match m_i to $\hat{\theta}$ item selection, a constant d is calculated for each item pool according to Equation 1.16. In the 1-PL item pool, $d = 0$. Define $m_i = b_i + d$. In GSSS, DSS, and ZSS (using SD weight), the quasi-match m_i to $\hat{\theta}$ item selection

selects the item not yet administered whose m_i value is next smaller (or next greater) to the $\hat{\theta}$ value, depending upon whether the obtained score x is smaller (or greater) than the expected score at $\hat{\theta}$; in the 1-PL or the modified 3-PL item pools, during the item selection process, the quasi-match m_i to $\hat{\theta}$ select item not yet administered whose m_i is closest to $\hat{\theta}$, in the direction where the MLE lies.

In the maximum information item selection method, after each item is administered and the current ability is estimated, the information of each item not yet administered at the current ability estimate is calculated. The item that has the highest information is chosen to present.

CAT strategies. In the study of comparison of the quasi-match m_i to $\hat{\theta}$ and the maximum information item selection methods, GSSS, DSS, and ZSS (using SD weight) were the same as in Study One. The SD weight used for the three CAT was 0.7.

In the study of comparison of the two versions of ZSS, the SD weight used in ZSS (using SD weight) was 0.4. Item selection methods were the quasi-match m_i to $\hat{\theta}$ item selection and the maximum information item selection. Other aspects of ZSS (using SD weight) were the same as in Study One. In ZSS (no SD weight), Z-score estimate was calculated according to Equation 1.26 (or Equation 1.28, since the locally best weight was applied) after each item was administered. No statistical hypothesis testing is involved in the current ability estimation process. After each item was admonistered, the current ability estimate was the Z-score estimate evaluated at the previous ability estimate. Next item was selected based on the current ability estimate and administered. A new Z-score estimate was calculated. The process continued

until the number of items administered was 20. In both versions of ZSS, during the current ability estimation process, if the current ability estimate was bigger than 3.4 (or smaller than -3.4), the current ability estimate was assigned to be 3.4 (or -3.4). The item selection method was the maximum information item selection.

Final ability estimate for all CAT strategies in Study Two was the MLE of ability. Dichotomous search was used to solve the MLE of ability. Iteration process continued until the difference between two successive estimated θ values was less than 0.001. In the final ability estimation, if a simulee's item responses were all 1s or all 0s, or a simulee's MLE solution was less than -10, or greater than 10, it would be excluded from the results. Each excluded case was replaced by a simulee with the same ability level, to maintain 100 simulees in each group.

Results

Comparison of two item selection methods

Table 3.1 shows the measurement precision and executing time for GSSS, DSS, and ZSS using SD weight 0.7, in the 1-PL and the modified 3-PL item pools. Value in each cell except for the last two columns, represents the average bias, absolute errors, MSE, and information for 2500 simulees grouped in 25 ability levels. The value in each cell of the next to the last column is the replacements made for each CAT strategy using each item selection method, in each of the item pools. The value in each cell of the last column is the executing time (seconds) of 2500 simulees for each CAT strategy.

Table 3.1: Measurement precision and executing time for GSSS, DSS, and ZSS using quasi-match m_i to $\hat{\theta}$ and maximum information item selection methods, in the 1-PL and the modified 3-PL item pools[a]

CAT strategy	Bias	ABE	MSE	Info	Replace	Time
1-PL						
GSSS						
Quasi[b]	-0.011	0.229	0.085	11.884	0	71"
Max info[c]	-0.011	0.228	0.084	11.831	0	180"
DSS						
Quasi	-0.012	0.234	0.086	11.885	0	68"
Max info	-0.011	0.233	0.086	11.830	0	177"
ZSS						
Quasi	-0.008	0.230	0.087	11.793	0	44"
Max info	-0.008	0.230	0.086	11.733	0	153"
Modified 3-PL						
GSSS						
Quasi	0.032	0.269	0.133	9.629	4	70"
Max info	0.028	0.268	0.132	9.581	2	179"
DSS						
Quasi	0.032	0.273	0.137	9.582	1	67"
Max info	0.029	0.273	0.133	9.549	1	177"
ZSS						
Quasi	0.034	0.272	0.137	9.512	4	45"
Max info	0.027	0.271	0.137	9.506	3	156"

[a]2500 simulees in each cell. SD weight = 0.7.

[b]Quasi-match m_i to $\hat{\theta}$ item selection.

[c]Maximum information item selection.

Bias and frequencies of replacement. From Table 3.1 one can see that no matter what item selection method was used, the average bias for GSSS, DSS, and ZSS were very small in the 1-PL item pool. The average bias for the three CAT strategies were also small in the modified 3-PL item pool, but were slightly bigger than those in the 1-PL item pool. No replacement was made in the 1-PL item pool for all three CAT strategies using SD weight 0.7. In the modified 3-PL item pool, a few simulees had negative infinity final MLE ability estimates. Those simulees were replaced by simulees with the same ability levels. The total frequencies of replacement in each CAT strategy were small.

Absolute errors, MSE, information, and executing time Table 3.1 also shows the average absolute errors, MSE, information, and executing time for GSSS, DSS, and ZSS. From Table 3.1 one can see that there was no significant difference between each pair of data for each CAT strategy that used the two different item selection methods, except for data of the last column, which was the executing time. The average bias, absolute errors, MSE, and information were almost the same by using the two item selection methods either in the 1-PL or the modified 3-PL item pools. From the last column one could see that the executing time for using the quasi-match m_i to $\hat{\theta}$ item selection was about two fifths that for using the maximum information item selection method in GSSS and DSS, and about two sevenths that for using the maximum information item selection method in ZSS, in both of the 1-PL and the modified 3-PL item pools.

Comparison of two versions of ZSS

Table 3.2 shows the measurement precision and executing time for ZSS using SD weight 0.4, using the quasi-match m_i to $\hat{\theta}$ and the maximum information item selection methods, and for ZSS (no SD weight) using the maximum information item selection, in the 1-PL and the modified 3-PL item pools. Value in each cell except for the last two columns, represents the average bias, absolute errors, MSE, and information for 2500 simulees grouped in 25 ability levels. The value in each cell of the next to the last column is the replacements made for each CAT strategy using each item selection method, in each of the item pool. The value in each cell of the last column is the executing time (seconds) of 2500 simulees for each CAT strategy.

Bias and frequencies of replacement. From Table 3.2 one can see that the average bias for both versions of ZSS was very small in the 1-PL item pool. The average bias for the two versions of ZSS was also small in the modified 3-PL item pool, but slightly bigger than those in the 1-PL item pool. No replacement was made in the 1-PL item pool for any CAT. In the modified 3-PL item pool, a few simulees had negative infinity final MLE ability estimates. Those simulees were replaced by simulees with the same ability levels. The total frequencies of replacement in each CAT strategy were small.

Absolute errors, MSE, information, and the executing time Table 3.2 also shows the average absolute errors, MSE, information, and executing time for the two versions of ZSS. From Table 3.2 one can see that the measurement precision of using the quasi-match m_i to $\hat{\theta}$ item selection and the maximum information item

Table 3.2: Measurement precision and executing time for two versions of ZSS in the 1-PL and the modified 3-PL item pools[a]

CAT strategy	Bias	ABE	MSE	Info	Replace	Time
1-PL						
Using SD weight 0.4						
Quasi[b]	-0.006	0.229	0.085	11.857	0	47"
Max info[c]	-0.009	0.228	0.084	11.859	0	160"
No SD weight						
Max info	-0.007	0.230	0.085	11.958	0	151"
Modified 3-PL						
Using SD weight 0.4						
Quasi	0.041	0.269	0.138	9.613	6	47"
Max info	0.035	0.271	0.137	9.700	5	159"
No SD weight						
Max info	0.037	0.272	0.139	9.703	4	152"

[a] 2500 simulees in each cell.

[b] Quasi-match m_i to $\hat{\theta}$ item selection.

[c] Maximum information item selection.

selection for ZSS using SD weight 0.4 was almost the same in the 1-PL and the modified 3-PL item pools. ZSS (using SD weight) could measure as precisely as ZSS (no SD weight) in the 1-PL or the modified 3-PL item pools. The average bias, absolute errors, MSE, and information were almost the same for both versions of ZSS in the 1-PL or the modified 3-PL item pools. The executing time for ZSS using SD weight and the quasi-match m_i to $\hat{\theta}$ item selection was about one third as that for ZSS using the maximum information item selection method, in both of the 1-PL and the modified 3-PL item pools. Using the maximum information item selection method, the executing time for ZSS (no SD weight) was only slightly less than that for ZSS (using SD weight).

Discussion

Results of Study Two showed that there was no difference in measurement precision between the two item selection methods—the quasi-match m_i to $\hat{\theta}$ item selection and the maximum information item selection, for GSSS, DSS, and ZSS, in the 1-PL and the 3-PL item pools. The executing time for the three CAT strategies using the quasi-match m_i to $\hat{\theta}$ item selection was much less than that of using the maximum information item selection. There was no difference in measurement accuracy between the two versions of ZSS—ZSS (using SD weight) and ZSS (no SD weight).

In the quasi-match m_i to $\hat{\theta}$ item selection, a constant d is calculated for each item pool according to Equation 1.16. Define $m_i = b_i + d$. If the current ability estimate is an interval estimate, such as in GSSS, DSS, and ZSS (using SD weight), the quasi-match m_i to $\hat{\theta}$ item selection selects the item not yet administered, whose

m_i value is next smaller (or next greater) to the $\hat{\theta}$ value, depending upon whether the obtained score is smaller (or greater) than the expected score at $\hat{\theta}$. In the 1-PL and the modified 3-PL item pools, the as and the cs are constant for all items in each item pool. After an item is administered and the current ability is determined, the item selected by the maximum information item selection is the same item selected by the match m_i to $\hat{\theta}$ item selection. That will be the same or slightly easier (or slightly more difficult) than that item if the quasi-match m_i to $\hat{\theta}$ item selection is used in GSSS, DSS, and ZSS. The maximum information item selection method is one of the most sophisticated item selection methods that can select the most informative items to administer in a CAT. In the 1-PL and the modified 3-PL item pools, GSSS, DSS, ZSS (using SD weight) can apply the quasi-match m_i to $\hat{\theta}$ item selection method to achieve the same measurement precision as does the maximum information item selection, and drastically decrease the computation burden.

The measurement precision of the two versions of ZSS—ZSS (using SD weight) and ZSS (no SD weight)—was almost the same in the 1-PL and the modified 3-PL item pools. In the 1-PL and the modified 3-PL item pools, ZSS using an appropriate SD weight, applying the quasi-match m_i to $\hat{\theta}$ item selection method can achieve the same measurement precision as in ZSS (no SD weight) using the maximum information item selection, and drastically reduce the executing time.

In the present study the comparison of the two versions of ZSS was limited in the 1-PL and the 3-PL item pools. The measurement precision was only checked for the grand means of bias, absolute errors, MSE, and information in each CAT strategy. It has been shown in Study One that the measurement precision of ZSS using different SD weights was slightly different. The measurement precision of ZSS (using an

appropriate SD weight) and ZSS (no SD weight) should not differ significantly in most situations.

ZSS (no SD weight) provides an updated Z-score estimate as the current ability estimate after each item is administered. In GSSS, DSS, and ZSS (using SD weight), the current ability estimate could either be the same as the previous ability estimate, or a new one, depending upon the hypothesis testing results. ZSS (no SD weight) always provides a new current ability estimate after each item is administered.

CHAPTER 4. STUDY THREE: MEASUREMENT PRECISION FOR GSSS, DSS, ZSS, AND MLES IN THE 1-PL AND THE MODIFIED 3-PL ITEM POOLS

Study Two showed that in the 1-PL and the 3-PL item pools there was no difference in measurement accuracy between the two item selection methods—the quasi-match m_i to $\hat{\theta}$ and the maximum information item selection, for GSSS, DSS, and ZSS. There was no difference in measurement accuracy between the two versions of ZSS—ZSS (using SD weight) and ZSS (no SD weight). The present study tried to compare the measurement accuracy and efficiency for GSSS, DSS, and ZSS (no SD weight), with MLES, using the maximum information item selection method in the 1-PL and the 3-PL item pools. In the item selection process, statistical hypothesis testing was used in GSSS, and DSS to determine the current ability estimate. The current ability estimate determined by GSSS and DSS is not necessarily equal to the MLE of ability, but is within a confidence interval of the MLE. During the item selection and the current ability estimation process, the Z-score estimate evaluated at the previous ability estimate is the current ability estimate in ZSS. The MLE is the current ability estimate in MLES. Study Three intends to compare measurement precision of GSSS, DSS, ZSS, and MLES, using the same maximum information item selection method and final MLE ability estimation procedure. It is expected that

GSSS, DSS, ZSS are more robust against any aberrant responses and provide more precise measurement results.

Design

Monte Carlo studies were conducted to compare the measurement precision of GSSS, DSS, ZSS, and MLES, in the 1-PL and the modified 3-PL item pools. In each item pool, the measurement accuracy and efficiency of the ability estimates of the four CAT strategies were compared. Each CAT contained 20 items. Two estimated precision indices—absolute errors, and test information—were used separately as dependent variables in Split Plot Factorial (SPF) 25 (ability level) x 4 (Strategy) ANOVA analysis. Two other estimated precision indices—bias and MSE—were also compared among the four CAT strategies. All CATs were conducted by Digital UNIX DEC Station 3100, using simulated examinees. The programs were developed by the author using the C language.

Method

Simulees and item pools

A simulee was a computer generated simulee with a true ability value θ. There were 2500 simulees in each of the four CAT strategies—GSSS, DSS, ZSS, and MLES in each item pool, with 100 as a group at each of 25 ability levels equally spaced in [-3, 3], in interval of 0.25.

The 1-PL item pool and the modified 3-PL item pool were the same as in Study One. Item response simulation was the same as in Study One.

CAT strategies

All simulees were assumed a pre-ability estimate $\theta = 0$. The MLE of ability was applied in the final ability estimation. In the final ability estimation, if a simulee's item responses were all 1s or all 0s, or a simulee's MLE solution was less than -10, or greater than 10, it would be excluded from the results. Each excluded case was replaced by a simulee with the same ability level. The item selection method for each CAT was the maximum information item selection. The final ability estimation was the MLE of ability. Each CAT contained 20 items.

GSSS and DSS. GSSS and DSS were the same as in Study One. *SD* weight 0.7 was used in determining the confidence interval of the expected score at each testing point.

ZSS. ZSS (no *SD* weight) used in Study Three was the same as in Study Two. Z-score estimate was calculated after each item was administered. No statistical hypothesis testing is involved in the item selection and current ability estimation process. Each Z-score estimate evaluated at the previous ability estimate after each item was administered was the current ability estimate. The next item was selected based on the current ability estimate and administered. A new Z-score estimate was calculated. The process continued until the number of items administered was 20. During the item selection process, if a simulee's current ability estimate was bigger than 3.4 (or smaller than -3.4), the simulee's current ability estimate was assigned to be 3.4 (or -3.4).

MLES. The maximum likelihood ability estimation was applied in MLES for the current ability estimation and the final ability estimation. The item selection algorithm used for the present research for MLES was similar to that used by Stocking (1987) and proceeded as follows: (a) Select the first item to be administered that has the maximum information at ability $\theta = 0$. (b) Select the second item that is maximally informative at an extremely low (high) ability level if the first item is answered incorrectly (correctly). If possible, compute MLE. (c) If it is impossible to compute MLE, continue to select subsequent items that are maximally informative at extreme ability levels until it is possible to compute MLE. The current ability estimate $\hat{\theta}$ is the MLE. After each item is administered, compute a new MLE, and the next item to be presented is the item that has not yet been presented that can maximize information at the MLE. (d) Test is terminated after 20 items are administered. The final ability estimate is the final MLE after 20 items are administered. The method to solve MLE is dichotomous search. Iteration process is continued until the difference between two successive estimated θ values is less than 0.001.

Results

In the 1-PL item pool

In the 1-PL item pool, there was difference among the measurement precision of GSSS, DSS, ZSS and MLES. Four precision indices—bias, absolute errors, MLE, and information—were compared, respectively, among the four strategies. No simulee was excluded from the results.

Bias. For each 100-simulee group, the average bias for GSSS, DSS, ZSS, and MLES was very small. (see Table 4.1). The range of the group average bias for GSSS, DSS, ZSS, and MLES, in order, was: (-0.075, 0.049), (-0.066, 0.049), (-0.057, 0.056), (-0.078, 0.056). From Table 4.2 one can see that the average bias for each CAT strategy was very small in the 1-PL item pool.

Absolute errors. Table 4.2 showed that in the 1-PL item pool the means of absolute errors for GSSS, ZSS, DSS, and MLES were: 0.228, 0.230, 0.233, and 0.233, respectively. Results of the SPF 25 (ability level) x 4 (Strategy) ANOVA analysis are shown in Table 4.3. Difference among the means of absolute errors for GSSS, ZSS, DSS, and MLES was not significant ($F(3, 7425) = 1.19$, $p > .05$). There was no significant difference among the means of absolute errors of 25 levels ($F(24, 2475) = 0.65$, $p > .05$). The interaction between the CAT strategies and ability levels was also not significant ($F(72, 7425) = 0.99$, $p > .05$).

MSE. The average MSEs of GSSS, DSS, ZSS, and MLES were: 0.084, 0.086, 0.085, and 0.088, respectively. Figure 4.1 showed that the MSEs for GSSS, DSS, ZSS, and MLES were not different from each other.

Test information. From Table 4.2 one can see that the means of test information for ZSS, DSS, GSSS, and MLES in the 1-PL item pool were: 11.958, 11.838, 11.831, and 11.551, respectively. Results of the SPF 25 (ability level) x 4 (Strategy) ANOVA are shown in Table 4.4. There was significant difference among the four means of test information ($F(3, 7425) = 103.89$, $p < .01$). Further comparison among the four means using LSD ($\alpha = 0.01$, $df = 7425$, $MSE = 0.7175$, $n = 2500$) showed

Table 4.1: Bias (θ) for GSSS, DSS, ZSS, and MLES at each ability level in the 1-PL item pool[a]

θ	GSSS	DSS	ZSS	MLES
-3.00	-0.037	-0.033	-0.019	-0.023
-2.75	-0.024	-0.016	-0.024	-0.023
-2.50	0.006	0.021	0.009	0.015
-2.25	0.011	0.024	-0.006	0.028
-2.00	-0.063	-0.058	-0.057	-0.078
-1.75	-0.031	-0.019	0.001	-0.008
-1.50	-0.000	0.009	-0.014	-0.007
-1.25	0.012	0.040	0.038	0.035
-1.00	-0.040	-0.036	-0.023	-0.030
-0.75	-0.006	-0.014	-0.022	-0.035
-0.50	0.043	0.036	0.056	0.056
-0.25	0.036	-0.005	0.031	0.047
0.00	-0.004	0.003	0.003	0.019
0.25	-0.031	-0.049	-0.037	-0.058
0.50	0.049	0.049	0.041	0.051
0.75	-0.044	-0.029	-0.042	0.001
1.00	-0.028	-0.055	-0.043	-0.062
1.25	-0.004	0.005	0.013	0.003
1.50	0.026	0.013	0.018	0.010
1.75	-0.002	-0.018	0.016	-0.014
2.00	-0.041	-0.025	-0.049	-0.040
2.25	-0.019	-0.013	-0.017	-0.047
2.50	-0.075	-0.066	-0.049	-0.051
2.75	0.040	0.013	0.026	0.013
3.00	-0.060	-0.065	-0.036	-0.071

[a]100 simulees in each cell.

Table 4.2: Measurement precision for GSSS, DSS, ZSS, and MLES in the 1-PL and the modified 3-PL item pools[a]

CAT strategy	Bias	ABE	MSE	Info	Replace
1-PL					
GSSS	-0.011	0.228	0.084	11.831	0
DSS	-0.011	0.233	0.086	11.838	0
ZSS	-0.007	0.230	0.085	11.958	0
MLES	-0.011	0.233	0.088	11.551	0
Modified 3-PL					
GSSS	0.028	0.268	0.132	9.581	2
DSS	0.029	0.273	0.133	9.549	1
ZSS	0.037	0.272	0.139	9.703	4
MLES	0.060	0.301	0.199	8.926	3

[a]2500 simulees in each cell.

that the means of information for ZSS, DSS, GSSS, and MLES were significantly different. Mean information obtained by MLES was significantly lower than that obtained by any of the other three strategies. The higher information was achieved by ZSS, which was significantly higher than those achieved by the other three CAT strategies. No significant difference was observed between the means of information for DSS and GSSS. There was significant difference among the means of test information of 25 ability levels ($F(24, 2475) = 2.81$, $p < .01$). Significant interaction was also observed between CAT strategies and ability levels ($F(72, 7425) = 8.54$, $p < .01$).

Figure 4.2 shows the information curves for GSSS, DSS, ZSS, and MLES. In the extremely lower and extremely higher levels of ability, information obtained by

Table 4.3: ANOVA table for SPF 25x4 design in the 1-PL item pool, with absolute errors as the dependent variable

Source	*SS*	*df*	*MS*	*F*
Between blocks				
L (Ability level)	1.3895	24	0.0579	0.65
S(L)	221.1677	2475	0.0894	
Within blocks				
B (CAT strategy)	0.0475	3	0.0158	1.19
BL	0.9534	72	0.0132	0.99
BS(L)	99.2152	7425	0.0134	
Total	322.7733	9999		

GSSS, DSS, and ZSS was almost the same. Information obtained by MLES was slightly higher than those of the others. However, in a wide range of other ability levels, information obtained by MLES was significantly lower than those of the other three strategies.

In summary, in the 1-PL item pool GSSS, DSS, and ZSS provided significantly more efficient ability estimates than did MLES. No significant difference in measurement accuracy among the four CAT strategies was found. ZSS was not significantly more precise than those of GSSS and DSS, but measured more efficiently than did GSSS and DSS.

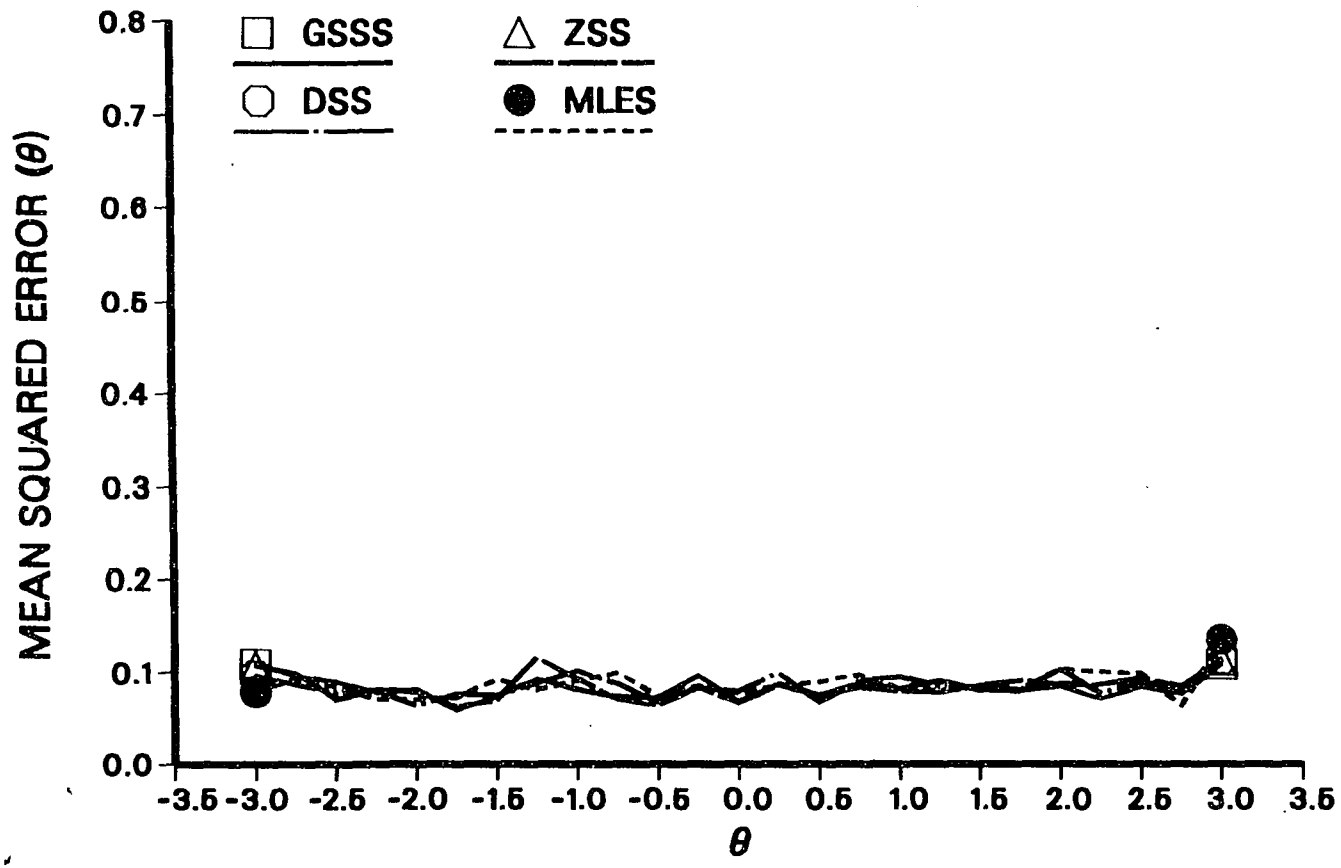

Figure 4.1: Mean squared errors for GSSS, DSS, ZSS, and MLES in the 1-PL item pool

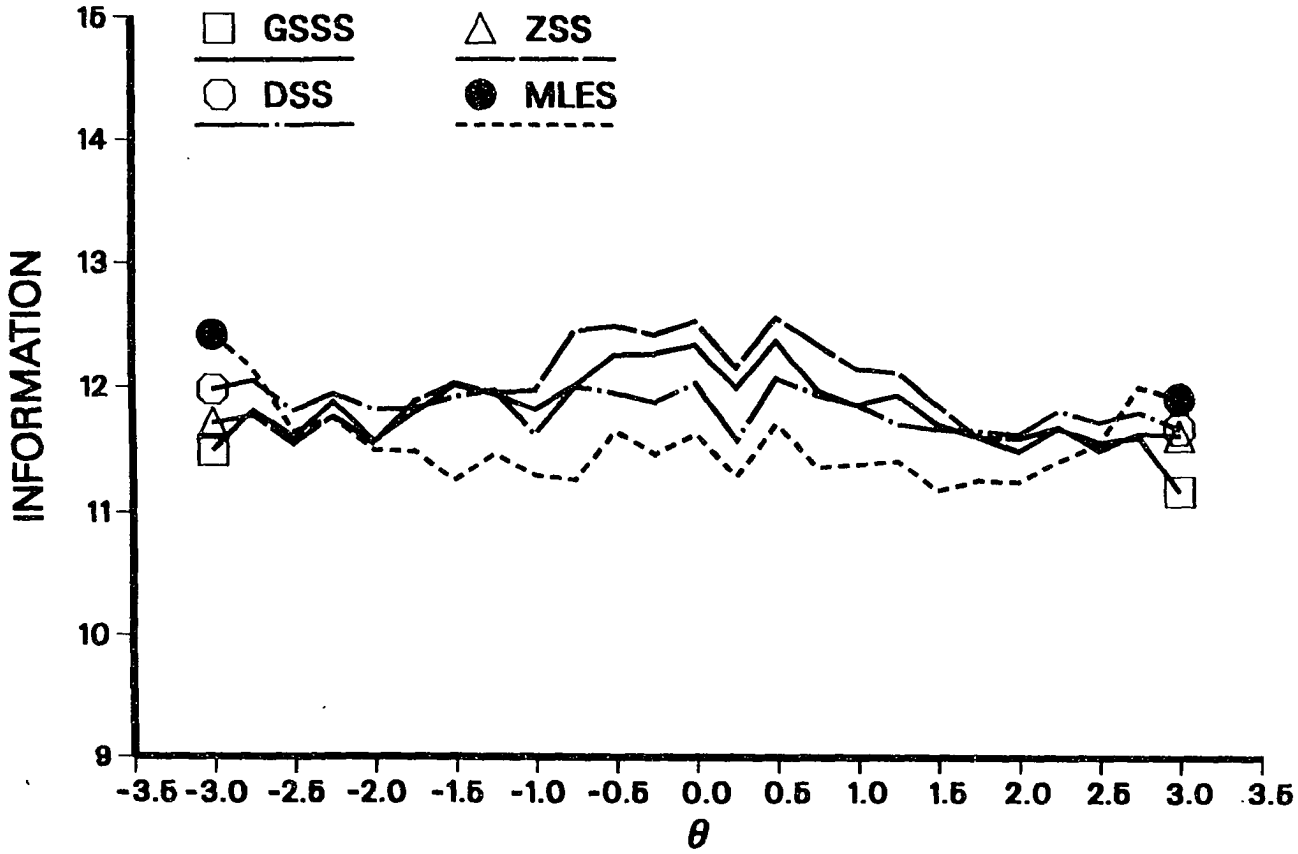

Figure 4.2: Test information for GSSS, DSS, ZSS, and MLES in the 1-PL item pool

Table 4.4: ANOVA table for SPF 25x4 design in the 1-PL item pool, with information as the dependent variable

Source	*SS*	*df*	*MS*	*F*
Between blocks				
L (Ability level)	331.9921	24	13.8330	2.81*
S(L)	12184.4688	2475	4.9230	
Within blocks				
B (CAT strategy)	223.6190	3	74.5397	103.89*
BL	441.3137	72	6.1294	8.54*
BS(L)	5327.2546	7425	0.7175	
Total	18508.6482	9999		

*$p < .01$

In the modified 3-PL item pool

In the modified 3-PL item pool, there was significant difference among the measurement precision of GSSS, DSS, ZSS and MLES. Four precision indices—bias, absolute errors, MSE, and information—were compared, respectively, among the four CAT strategies. Due to guessing effect, a few simulees had aberrant responses. Negative infinity ability estimates occasionally happened. Those simulees were replaced by simulees with the same ability levels.

Bias and frequencies of replacement. For each 100-simulee group, the average bias for GSSS, DSS, ZSS, and MLES was small, but greater than those in

the 1-PL item pool (see Table 4.5). The range of the group average bias for GSSS, DSS, ZSS, and MLES, in order, was: (-0.041, 0.142), (-0.033, 0.114), (-0.037, 0.140), and (-0.039, 0.148). Table 4.2 shows the average bias for each CAT in the modified 3-PL model. The bias for the four CAT strategies was also small in the modified 3-PL model, but was slightly bigger than those in the 1-PL model. Frequencies of replacement were small.

Absolute errors. Table 4.2 showed that the means of absolute errors for GSSS, ZSS, DSS, and MLES were: 0.268, 0.272, 0.273, and 0.301, respectively. Results of the SPF-25 (ability level) x 4 (Strategy) ANOVA analysis are shown in Table 4.6. Difference among the means of absolute errors for GSSS, ZSS, DSS, and MLES was significant ($F(3, 7425) = 9.74$, $p < .01$). Further comparison among the four means using LSD ($\alpha = 0.01$, $df = 7425$, $MSE = 0.0603$, $n = 2500$) showed that the mean absolute errors for MLES was significantly greater than that for GSSS, ZSS, or DSS. No significant difference was found among the three means of absolute errors for GSSS, ZSS, and DSS. There was significant difference among the means of absolute errors of 25 levels ($F(24, 2475) = 2.93$, $p < .01$). The interaction between CAT strategies and ability levels was not significant ($F(72, 7425) = 0.81$, $p > .05$).

MSE. Table 4.2 showed that the average MSEs for GSSS, DSS, ZSS, and MLES in the modified 3-PL item pool were: 0.132, 0.133, 0.139, and 0.199, respectively. Figure 4.3 showed that the MSEs for MLES were significantly greater than those for GSSS, DSS, and ZSS, especially in the lower and middle range of the ability levels. The amount of MSEs for GSSS, DSS, and ZSS were almost the same.

Table 4.5: Bias (θ) for GSSS, DSS, ZSS, and MLES at each ability level in the modified 3-PL item pool[a]

θ	GSSS	DSS	ZSS	MLES
-3.00	0.042	0.073	0.059	0.146
-2.75	0.013	-0.012	0.010	0.029
-2.50	0.042	0.037	0.078	0.208
-2.25	0.059	0.087	0.128	0.137
-2.00	-0.010	-0.015	-0.023	0.059
-1.75	0.095	0.045	0.110	0.104
-1.50	0.065	0.012	0.018	0.132
-1.25	0.057	0.067	0.038	0.049
-1.00	-0.007	-0.006	0.029	0.059
-0.75	0.014	0.021	0.037	0.039
-0.50	0.142	0.114	0.140	0.148
-0.25	0.086	0.094	0.088	0.078
0.00	0.002	0.021	0.037	0.061
0.25	-0.005	0.003	-0.011	0.039
0.50	0.084	0.059	0.093	0.081
0.75	0.031	0.028	0.003	0.011
1.00	-0.031	-0.033	-0.004	-0.018
1.25	0.019	0.018	0.014	0.084
1.50	0.013	0.057	0.044	0.029
1.75	0.005	0.046	0.037	0.026
2.00	0.006	0.027	0.007	0.000
2.25	-0.028	-0.028	-0.003	-0.002
2.50	0.005	0.010	-0.019	-0.025
2.75	0.048	0.017	0.046	0.051
3.00	-0.041	-0.021	-0.037	-0.039

[a]100 simulees in each cell.

Table 4.6: ANOVA table for SPF 25x4 design in the modified 3-PL item pool, with absolute errors as the dependent variable

Source	*SS*	*df*	*MS*	*F*
Between blocks				
L (Ability level)	7.7471	24	0.3228	2.93*
S(L)	272.6691	2475	0.1102	
Within blocks				
B (CAT strategy)	1.7611	3	0.5870	9.74*
BL	3.5278	72	0.0490	0.81
BS(L)	447.6366	7425	0.0603	
Total	733.3418	9999		

*$p < .01$

Test information. From Table 4.2 one could see that the means of test information for ZSS, GSSS, DSS, and MLES in the modified 3-PL item pool were: 9.703, 9.581, 9.549, and 8.926, respectively. Results of the SPF-25 (ability level) x 4 (Strategy) ANOVA are shown in Table 4.7. There were significant differences among the four means of test information ($F(3, 7425) = 62.27$, $p < .01$). Further comparison among the four means using LSD ($\alpha = 0.01$, $df = 7425$, $MSE = 4.8853$, $n = 2500$) showed that the mean information for ZSS, GSSS, or DSS, was significantly higher than that for MLES. No significant difference of test information was found among ZSS, GSSS, and DSS. There was significant difference among the means of test information for 25 ability levels ($F(24, 2475) = 9.49$, $p < .01$). Significant interaction

Table 4.7: ANOVA table for SPF 25x4 design in the modified 3-PL item pool, with information as the dependent variable

Source	*SS*	*df*	*MS*	*F*
Between blocks				
L (Ability level)	2008.3650	24	83.6819	9.49*
S(L)	21833.4602	2475	8.8216	
Within blocks				
B (CAT strategy)	912.6614	3	304.2205	62.27*
BL	1398.3277	72	19.4212	3.98*
BS(L)	36273.6549	7425	4.8853	
Total	62426.4692	9999		

*$p < .01$

was also observed between CAT strategies and ability levels ($F(72, 7425) = 3.98$, $p < .01$).

Figure 4.4 shows the information curves for GSSS, DSS, ZSS, and MLES. Information for GSSS, DSS, and ZSS was significantly higher than that for MLES in all ability levels except for the extremely higher and extremely lower levels of ability, where the MLES achieved slightly higher information than the other three strategies.

In summary, in the modified 3-PL item pool, GSSS, DSS, and ZSS provided significantly more accurate and efficient ability estimates than did MLES. No significant difference in measurement precision was found among GSSS, DSS, and ZSS.

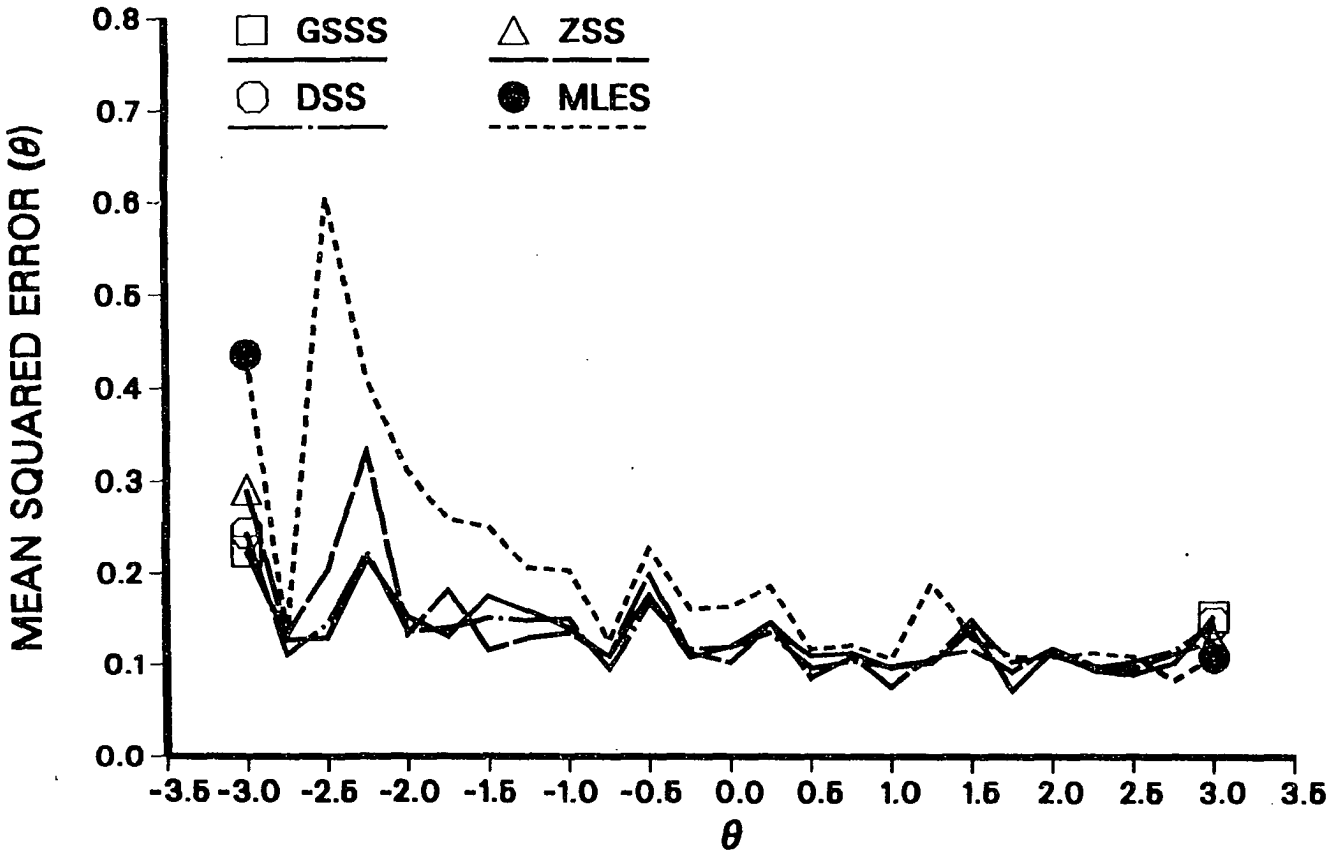

Figure 4.3: Mean squared errors for GSSS, DSS, ZSS, and MLES in the modified 3-PL item pool

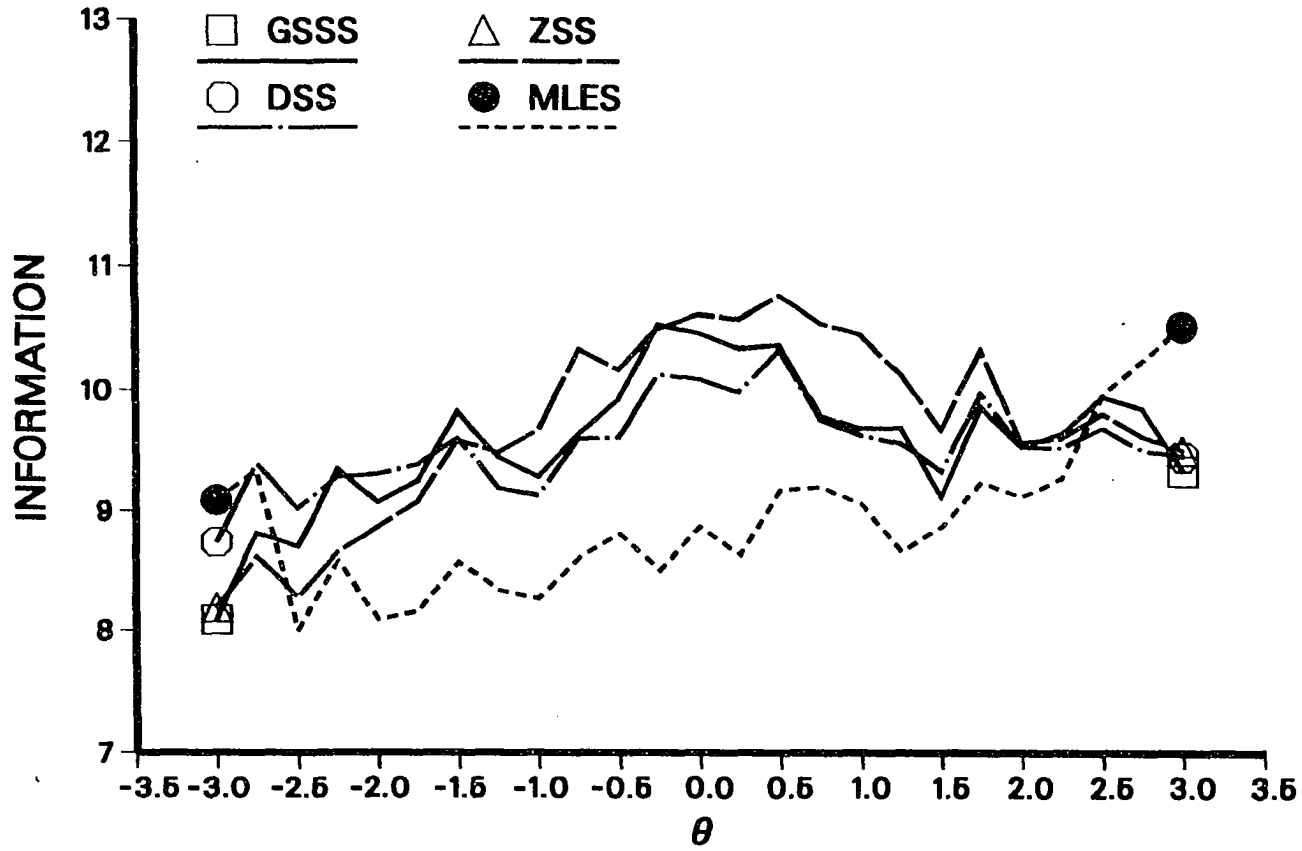

Figure 4.4: Test information for GSSS, DSS, ZSS, and MLES in the modified 3-PL item pool

Discussion

Results of Study Three showed that in the 1-PL item pool, GSSS, DSS, and ZSS provided more efficient but not more accurate ability estimates than did MLES. In the modified 3-PL item pool, where guessing effect existed, GSSS, DSS, and ZSS provided significantly more accurate and more efficient ability estimates than did MLES.

In the 1-PL item pool, there was no guessing effect. MLES measured almost as accurately as did GSSS, DSS, and ZSS, but less efficiently than did GSSS, DSS, and ZSS.

In the modified 3-PL item pool, GSSS, DSS, and ZSS measured much more precisely than did MLES. MLES produced much greater MSE than those of GSSS, DSS, and ZSS, had much lower information than those of GSSS, DSS, and ZSS, except in the extremely higher and extremely lower ability range.

Results of the present research showed robust and precise nature of GSSS, DSS, and ZSS. In GSSS and DSS, in determining each current ability estimate, a confidence interval of expected score at a testing point is used, to evaluate whether the obtained score is equal to the expected score at that testing point. In this way, the current ability estimate tends to remain in the previous estimate position, which is carefully chosen according to some optimal search methods. An examinee has a chance to recover from previous aberrant item responses. An examinee who makes a few lucky guessings during the earlier stage of a CAT will not get very high current ability estimate. In GSSS and DSS, adjustment is made for an obtained score that is lower than a score contributed by guessing during the item selection process. It prevents an examinee from getting a very low current ability estimate during the earlier stage

of a CAT. Thus an examinee who missed a few items by mistake in the earlier stage of a CAT will have a substantial chance to achieve an ability estimate that would be appropriate to him/her.

In GSSS and DSS, after each item is administered, successive hypothesis testing is conducted, starting from the testing point of the original search region, until a current estimate is determined. The systematical search process enables GSSS and DSS to find a current ability estimate at which the obtained score is within a confidence interval of the expected score, and the current ability estimate happens to be one of the successive testing points of golden section search or dichotomous search regions. The only difference between GSSS and DSS is the ratio of the sizes of the successive search regions. In GSSS, search region reduces by a ratio approximated to 0.618. In DSS, search region reduces by a ratio 0.5. Though the overall measurement precision of GSSS and DSS was not significantly different in Study Three, the results of Study Three indicated that generally, in the middle range of ability, information of GSSS was slightly higher than that of DSS, while in the extremely lower and higher ability levels, information of DSS was slightly higher than that of GSSS.

In ZSS, no search region is defined. The current ability estimate is the Z-score estimate. The Z-score estimate is evaluated at the previous ability estimate. Thus it is conditioned on the previous ability estimate. If the test information at the previous ability estimate is very low, the Z-score estimate might be far apart from the true ability level. It is important for ZSS to have informative items available at each current ability estimate. Otherwise, the measurement precision of ZSS may be affected. Generally, as the number of items administered increases, the Z-score estimate is approaching the MLE of ability. Adjustment is made for an obtained

score that is lower than a score contributed by guessing during the item selection process in ZSS.

In Study Three, the overall measurement efficiency of ZSS was significantly higher than those of GSSS, DSS and MLES in the 1-PL item pool, was significantly higher than that of MLES and was the same as those of GSSS and DSS in the modified 3-PL item pool. The information obtained by ZSS was slightly higher than those obtained by GSSS and DSS in a small range of the middle ability levels. The information for ZSS was slightly lower than those for GSSS and DSS in the extremely lower or extremely higher levels of ability levels. MLES measured slightly more precisely than GSSS, DSS, and ZSS only at the extremely low or extremely higher levels of ability in both of the item pools. This was partially due to the arbitrary predetermined item selection sequences when there was no finite MLE solution. In the present study, in the item selection process of MLES, if an examinee's item responses were all 1s (or all 0s), an item that was maximally informative at an extremely low (high) ability level was selected to present. Therefore, if an examinee's ability level was extremely low (or high), after the first item was administered, an item that could maximize information at the extremely low (or high) level was presented, providing that the examinee's response to the first item was in the right direction. That is, higher (lower) ability examinees responded correctly (incorrectly) to the first item. In the modified 3-PL item pool, MLES provided more precise ability estimates than the other three CAT strategies in the extremely higher levels of ability. This was also partially due to the fact that an MLE of a higher ability examinee was less affected by random guessing than that of a lower ability examinee.

CHAPTER 5. STUDY FOUR: MEASUREMENT PRECISION OF GSSS, DSS, ZSS, AND MLES USING THREE ITEM SELECTION METHODS, IN THE HYPOTHETICAL 3-PL POOL

In Studies One, Two, and Three, either in the 1-PL item pool or in the modified 3-PL item pool, the a_i and c_i parameters were constants. In those situations, in the item selection process, the maximum information item selection selects an item that is identical to the item if the match m_i to $\hat{\theta}$, or the quasi-match m_i to $\hat{\theta}$ item selection methods is used, when the current ability estimate is a point estimate (such as those in MLES and ZSS (noSD weight)). The quasi-match m_i to $\hat{\theta}$ selects item that is identical or slightly easier (or slightly more difficult) than the item selected by the maximum information item selection method, when the current ability estimate is an interval estimate (such as those in GSSS, DSS, and ZSS (using SD weight)).

In Studies One, Two, and Three, only the 1-PL and the modified 3-PL item pools were used. In real test situation, item pools containing items with various a_i values and various c_i values are common. A 3-PL model may be more suitable for CAT purpose. Wainer and Mislevy (1990) pointed out that the 3-PL was the most commonly applied IRT model in large scale testing applications. More sophisticated item selection methods, such as the maximum information item selection, could be used to achieve more accurate and more efficient measurement results.

Design

Monte Carlo studies were conducted to compare measurement precision of GSSS, DSS, ZSS (using SD weight), and MLES, in a hypothetical 3-PL item pool, using three item selection methods—the quasi-match m_i to $\hat{\theta}$ item selection, the match m_i to $\hat{\theta}$ item selection, and the maximum information item selection. Measurement accuracy and efficiency of ability estimates using different CAT strategies and different item selection methods were compared. Each CAT contained 20 items. Two estimated precision indices—absolute errors, and test information—were used separately as dependent variables in SPF 25 (ability level) x 4 (CAT Strategy) x 3 (Item Selection Method) ANOVA analysis. Two other estimated precision indices—bias, and MSE in different CAT strategies—were compared, respectively, using each item selection method. All CAT was conducted by ZENITH 386/20 microcomputers using simulated examinees. Each computer was equipped with a 80387 math co-processor. Programs were developed by the author using GWBASIC language, then compiled into machine language. The computational efficiency for the four CAT strategies was compared by comparison of their executing time.

Method

Simulees

A simulee was a computer generated simulee with a true ability value θ. There would be 2500 simulees in each of the four CAT strategies—GSSS, DSS, ZSS, and MLES—using each of the three item selection methods—the quasi-match m_i to $\hat{\theta}$, the match m_i to $\hat{\theta}$, and the maximum information item selections, with 100 as a

group at each of 25 ability levels equally spaced in [-3, 3], in interval of 0.25. Every numerical value was calculated during the running of the programs. No *info table* or other previous calculated tabulated values were used.

Item pool

An hypothetical 3-PL item pool was generated by a SAS program (see Appendix A). 200 items whose a_is were normally distributed with mean 1 and SD 0.2, b_is uniformly distributed across a range of -3.4 to 3.4 logits, and c_is normally distributed with mean 0.20 and SD 0.03 were generated. Urry (1977) suggested that in a CAT item pool the as of items should be at least equal to 0.8. The bs should be uniformly distributed. The cs should be less than 0.3. There were 33 items whose a_i parameters were smaller than 0.8, which were excluded from the item pool. The 3-PL hypothetical item pool in the present research contained the remaining 167 items (see Appendix C). The mean and the standard deviation were 1.063 and 0.160 for as, -0.052 and 2.012 for bs, 0.198 and 0.028 for cs, respectively.

CAT strategies

Three item selection methods—the quasi-match m_i to $\hat{\theta}$ item selection, the match m_i to $\hat{\theta}$ item selection, and the maximum information item selection—were used in each CAT strategy. The MLE of ability was used in the final ability estimation for each CAT strategy. Every simulee's ability level was assumed to be $\theta = 0$ before testing. In final ability estimation, if a simulee's item responses were all 1s or all 0s, or a simulee's MLE solution was less than -10, or greater than 10, it would be excluded from the results. Each excluded case was replaced by a simulee with the same ability

level. In GSSS and DSS, the range of the original search region was [-3.4, 3.4], which limited the current ability estimate into that range. In ZSS, in which there was no search region, the current ability estimate was limited to $[b_1, b_{167} + 0.25]$, where Item 1 and Item 167 were the easiest and most difficult items in the hypothetical item pool, respectively. In the item pool, that range was [-3.355, 3.631].

GSSS, and DSS were the same as in Study One. ZSS (using SD weight) was the same as in Study One. Statistical hypothesis testing was involved in determining the current ability estimate in GSSS, DSS, and ZSS. SD weight 0.7 was used in determining the confidence interval of the expected score at each testing point in GSSS, DSS, and ZSS. The final ability estimate in the three CAT strategies was the MLE of ability which was solved by the dichotomous search method.

MLES. In MLES, every simulee's ability level was assumed to be $\theta = 0$ before testing. Three item selection methods were applied in MLES in Study Four. During item selection process, the MLE of ability was the current ability estimate. If a simulee's item responses were all 1s (or all 0s), an item that can maximize information at the extreme high (or low) levels of ability was selected to present. The quasi-match m_i to $\hat{\theta}$ item selection, the match m_i to $\hat{\theta}$ item selection, and the maximum information item selection method were used in conjunction with the MLE ability estimation. Dichotomous search was used to solve the MLE of ability.

Results

In the hypothetical 3-PL item pool, there was significant difference among the measurement precision of GSSS, DSS, ZSS and MLES. Four precision indices—bias,

absolute errors, MSE, and test information—were compared, respectively, among the four strategies. There were a few simulees whose final ability estimates were negative infinity. They were replaced by simulees with the same ability levels.

Bias and frequencies of replacement. For each 100-simulee group, the average bias for GSSS, DSS, ZSS, and MLES using each of the three item selection methods, was generally small (see Tables 5.1, 5.2, and 5.3). The range of the group average bias for GSSS, DSS, ZSS, and MLES, in order, was: (-0.035, 0.151), (-0.035, 0.132), (-0.025, 0.150), and (-0.010, 0.201) in the quasi-match m_i to $\hat{\theta}$ item selection; (-0.032, 0.153), (-0.051, 0.157), (-0.062, 0.150), and (-0.025, 0.141) in the match m_i to $\hat{\theta}$ item selection; (-0.021, 0.080), (-0.040, 0.094), (-0.019, 0.122), and (-0.016, 0.137) in the maximum information item selection. From Table 5.4 one can see that the average bias for each CAT strategy using each item selection method was small. Using the quasi-match m_i to $\hat{\theta}$ item selection, only one or two replacements happened in GSSS, DSS, or ZSS. MLES had more replacements than those three strategies. Using the maximum information item selection, ZSS and MLES had more replacements than those in GSSS and DSS. However, when the match m_i to $\hat{\theta}$ item selection was used, no replacement was made in MLES. Frequencies of replacement were slightly higher in GSSS, DSS, and ZSS.

Absolute errors. Table 5.4 also shows the average absolute errors by CAT strategies and item selection methods. The overall means of absolute errors for GSSS, ZSS, DSS, and MLES were: 0.273, 0.279, 0.281, and 0.289, respectively. Results of the SPF 25 (ability level) x 4 (Strategy) x 3 (Item Selection Method) ANOVA analysis are shown in Table 5.5. Difference among the means of absolute errors for GSSS, ZSS,

Table 5.1: Bias (θ) for GSSS, DSS, ZSS, and MLES at each ability level using the quasi-match m_i to $\hat{\theta}$ item selection, in the hypothetical 3-PL item pool[a]

θ	GSSS	DSS	ZSS	MLES
-3.00	0.036	0.091	0.097	0.142
-2.75	0.151	0.097	0.150	0.201
-2.50	0.021	0.003	0.036	0.045
-2.25	0.093	0.132	0.062	0.158
-2.00	0.017	-0.003	0.056	0.043
-1.75	0.095	0.140	0.075	0.104
-1.50	0.030	0.045	0.072	0.026
-1.25	0.056	-0.002	0.037	0.051
-1.00	0.012	0.023	-0.003	0.004
-0.75	0.026	0.038	0.026	0.067
-0.50	0.040	0.016	0.045	0.102
-0.25	0.046	0.029	0.015	0.043
0.00	-0.018	-0.009	-0.019	0.007
0.25	0.120	0.079	0.101	0.098
0.50	0.048	0.016	0.017	-0.003
0.75	0.084	0.004	0.049	0.026
1.00	0.012	0.029	0.013	0.023
1.25	-0.015	-0.035	-0.002	0.009
1.50	0.031	0.018	0.054	0.064
1.75	0.019	-0.009	0.005	-0.010
2.00	0.043	0.024	0.050	0.016
2.25	-0.017	-0.012	0.028	0.015
2.50	0.017	0.014	0.013	0.023
2.75	-0.035	-0.006	-0.011	0.012
3.00	-0.004	0.010	-0.025	-0.009

[a]100 simulees in each cell.

Table 5.2: Bias (θ) for GSSS, DSS, ZSS, and MLES at each ability level using the match m_i to $\hat{\theta}$ item selection, in the hypothetical 3-PL item pool[a]

θ	GSSS	DSS	ZSS	MLES
-3.00	0.090	0.114	0.136	0.097
-2.75	0.153	0.157	0.150	0.141
-2.50	0.039	0.006	0.048	0.027
-2.25	0.043	0.097	0.077	0.122
-2.00	0.059	0.035	0.050	0.096
-1.75	0.090	0.085	0.085	0.125
-1.50	0.012	0.069	0.058	0.043
-1.25	0.011	0.034	-0.005	0.080
-1.00	0.038	-0.018	-0.010	-0.012
-0.75	0.006	0.031	0.013	0.094
-0.50	0.055	0.037	0.044	0.072
-0.25	0.013	0.024	0.051	0.047
0.00	-0.030	-0.014	-0.062	0.012
0.25	0.065	0.071	0.089	0.085
0.50	0.028	0.032	0.033	0.008
0.75	0.042	0.017	0.046	0.046
1.00	0.008	0.026	0.025	0.038
1.25	-0.032	-0.030	0.027	0.010
1.50	0.050	0.034	0.022	0.035
1.75	0.010	0.014	0.010	0.020
2.00	-0.005	0.023	0.071	0.017
2.25	0.004	-0.051	-0.032	0.016
2.50	0.015	0.036	0.010	0.037
2.75	-0.005	0.005	-0.004	0.006
3.00	0.012	0.015	-0.021	-0.025

[a]100 simulees in each cell.

Table 5.3: Bias (θ) for GSSS, DSS, ZSS, and MLES at each ability level using the maximum information item selection, in the hypothetical 3-PL item pool[a]

θ	GSSS	DSS	ZSS	MLES
-3.00	0.064	0.089	0.112	0.088
-2.75	0.079	0.051	0.122	0.072
-2.50	0.020	0.027	0.016	0.005
-2.25	0.080	0.086	0.106	0.137
-2.00	0.053	0.029	0.104	0.051
-1.75	0.078	0.094	0.073	0.124
-1.50	0.033	0.030	0.045	0.057
-1.25	-0.021	-0.015	0.043	0.035
-1.00	0.010	0.018	0.009	0.007
-0.75	-0.001	0.033	0.026	0.045
-0.50	0.028	0.024	0.036	0.071
-0.25	0.027	0.025	0.037	0.048
0.00	-0.018	-0.011	-0.019	-0.015
0.25	0.055	0.053	0.054	0.050
0.50	0.012	-0.007	0.034	0.012
0.75	-0.002	0.016	0.034	0.048
1.00	0.025	0.010	0.028	0.013
1.25	-0.018	-0.040	0.008	-0.011
1.50	0.036	0.044	0.023	0.071
1.75	0.008	0.013	-0.023	0.005
2.00	0.051	0.049	0.067	0.030
2.25	0.008	-0.016	0.003	0.013
2.50	0.018	0.018	0.018	0.056
2.75	-0.019	-0.009	0.006	-0.003
3.00	0.015	0.005	0.011	-0.016

[a]100 simulees in each cell.

Table 5.4: Measurement precision for GSSS, DSS, ZSS, and MLES in the hypothetical 3-PL item pool[a]

CAT strategy	Bias	ABE	MSE	Info	Replace
Quasi-Match m_i to $\hat{\theta}$					
GSSS	0.036	0.287	0.151	8.267	2
DSS	0.029	0.292	0.153	8.186	1
ZSS	0.038	0.285	0.147	8.303	2
MLES	0.050	0.300	0.187	7.769	6
Match m_i to $\hat{\theta}$					
GSSS	0.029	0.278	0.137	8.388	6
DSS	0.034	0.286	0.148	8.323	6
ZSS	0.036	0.285	0.144	8.347	9
MLES	0.048	0.297	0.173	7.814	0
Maximum Information					
GSSS	0.025	0.253	0.110	9.699	2
DSS	0.025	0.266	0.123	9.554	2
ZSS	0.039	0.266	0.123	9.523	6
MLES	0.040	0.269	0.140	9.206	5

[a]2500 simulees in each cell.

DSS, and MLES was significant ($F(3, 7425) = 5.83$, $p < .01$). Further comparison among the four means using LSD ($\alpha = 0.01$, $df = 7425$, $MSE = 0.0575$, $n = 7500$) showed that the mean absolute error for MLES was significantly greater than those for GSSS and ZSS, but was not significantly greater than that for DSS. The means of absolute errors for GSSS, ZSS, and DSS were not significantly different. There was significant difference among the means of absolute errors for 25 ability levels ($F(24, 2475) = 5.99$, $p < .01$). The interaction between CAT strategies and ability levels was not significant ($F(72, 7425) = 0.87$, $p > .05$).

The overall means of absolute errors using the three item selection methods—the quasi-match m_i to $\hat{\theta}$, the match m_i to $\hat{\theta}$, and the maximum information item selection—were: 0.291, 0.287, and 0.263, respectively. Difference among the means of absolute errors using the three item selection methods was significant ($F(2, 4950 = 36.25$, $p < .01$). Further comparison among the means of the three item selection methods using LSD ($\alpha = 0.01$, $df = 4950$, $MSE = 0.0617$, $n = 10000$) showed that the mean absolute error for the maximum information item selection was significantly smaller than those of the other item selection methods. There was no significant difference between measurement accuracy of the quasi-match m_i to $\hat{\theta}$ item selection, and the match m_i to $\hat{\theta}$ item selection. The interaction between item selection methods and ability levels was not significant ($F(48, 4950) = 0.86$, $p > .05$). The interaction between the CAT strategies and item selection methods was not significant ($F(6, 14850) = 0.54$, $p > .05$). The interaction among the CAT strategies, item selection methods and ability levels was also not significant ($F(144, 14850) = 0.48$, $p > .05$).

Table 5.5: ANOVA table for SPF 25x4x3 design in the 3-PL model, with absolute errors as the dependent variable

Source	*SS*	*df*	*MS*	*F*
Between blocks				
L (Ability level)	19.2231	24	0.8010	5.99*
S(L)	331.0051	2475	0.1337	
Within blocks				
B (CAT strategy)	1.0065	3	0.3355	5.83*
BL	3.6135	72	0.0502	0.87
BS(L)	426.6821	7425	0.0575	
C (Method)	4.4729	2	2.2365	36.25*
CL	2.5531	48	0.0532	0.86
CS(L)	305.1776	4950	0.0617	
BC	0.1934	6	0.0322	0.54
BCL	4.0998	144	0.0285	0.48
BCS(L)	885.9612	14850	0.0597	
Total	1983.9884	29999		

*$p < .01$

Overall test information. Table 5.4 also shows the average test information by CAT strategies and item selection methods. The overall means of test information for GSSS, ZSS, DSS, and MLES were: 8.785, 8.724, 8.688, and 8.263, respectively. Results of the SPF 25 (ability level) x 4 (Strategy) x 3 (Item Selection Method) ANOVA analysis are shown in Table 5.6. Difference among the test information for GSSS, ZSS, DSS, and MLES was significant ($F(3, 7425) = 160.66$, $p < .01$). Further comparison among the four means using LSD ($\alpha = 0.01$, $df = 7425$, $MSE = 2.6440$, $n = 7500$) showed that the means of information for GSSS, ZSS, and DSS were significantly higher than that of the MLES. There was no significant difference between means of information for GSSS and ZSS. There was no significant difference between means of information for ZSS and DSS. Information of GSSS was significantly higher than that of DSS. There was significant difference among the overall test information of 25 ability levels ($F(24, 2475) = 44.64$, $p < .01$). The interaction between CAT strategies and ability levels was significant ($F(72, 7425) = 9.33$, $p < .01$).

The overall test information using the three item selection methods—the quasi-match m_i to $\hat{\theta}$, the match m_i to $\hat{\theta}$, and the maximum information—were 8.131, 8.218, and 9.496, respectively. Difference among the test information using the three item selection methods was significant ($F(2, 4950) = 1810.22$, $p < .01$). Further comparison among the means for the three item selection methods using LSD ($\alpha = 0.01$, $df = 4950$, $MSE = 3.2234$, $n = 10000$) showed that the test information for the maximum information item selection was significantly higher than those of the other item selection methods. Information for the match m_i to $\hat{\theta}$ item selection was significantly higher than that for the quasi-match m_i to $\hat{\theta}$ item selection. The interaction between the item selection methods and ability levels was significant

($F(48, 4950) = 6.49$, $p < .01$). The interaction between the CAT strategies and item selection methods was significant ($F(6, 14850) = 3.02$, $p < .01$). The interaction among the CAT strategies, item selection methods and ability levels was also significant ($F(144, 14850) = 1.86$, $p < .01$).

MSE and information using the quasi-match m_i to $\hat{\theta}$ item selection. Using the quasi-match m_i to $\hat{\theta}$ item selection, the MSEs for GSSS, DSS, ZSS, and MLES were: 0.151, 0.153, 0.147, and 0.187, respectively. Figure 5.1 shows that the MSE for MLES is greater than those for GSSS, DSS, and ZSS, especially in the lower range of ability levels.

The means of test information for ZSS, GSSS, DSS, and MLES using the quasi-match m_i to $\hat{\theta}$ item selection were: 8.3033, 8.2667, 8.1862, and 7.7690, respectively. Information curves are shown in Figure 5.2. The highest information in the middle range of ability was achieved by ZSS. GSSS and DSS provided slightly lower information than ZSS in the middle range of ability, and slightly higher information in the extremely lower and higher levels of ability. Information obtained by MLES was significantly lower than those of the other CAT strategies, along the whole ability continuum, except for extremely higher levels of ability.

MSE and information using the match m_i to $\hat{\theta}$ item selection. Using the match m_i to $\hat{\theta}$ item selection the MSEs for GSSS, DSS, ZSS, and MLES were: 0.137, 0.148, 0.144, and 0.173, respectively. Figure 5.3 shows that the MSE for MLES is greater than those for GSSS, DSS, ZSS, especially in the lower range of ability levels.

The means of test information for GSSS, ZSS, DSS, and MLES using the match

Table 5.6: ANOVA table for SPF 25x4x3 design in the 3-PL model, with information as the dependent variable

Source	*SS*	*df*	*MS*	*F*
Between blocks				
L (Ability level)	8099.4005	24	337.4750	44.64*
S(L)	18708.8728	2475	7.5591	
Within blocks				
B (CAT strategy)	1274.3590	3	424.7863	160.66*
BL	1776.6540	72	24.6758	9.33*
BS(L)	19631.9586	7425	2.6440	
C (Method)	11670.1447	2	5835.0723	1810.22*
CL	1003.7566	48	20.9116	6.49*
CS(L)	15955.6333	4950	3.2234	
BC	53.9710	6	8.9952	3.02*
BCL	797.3840	144	5.5374	1.86*
BCS(L)	44210.0464	14850	2.9771	
Total	123182.1810	29999		

*$p < .01$

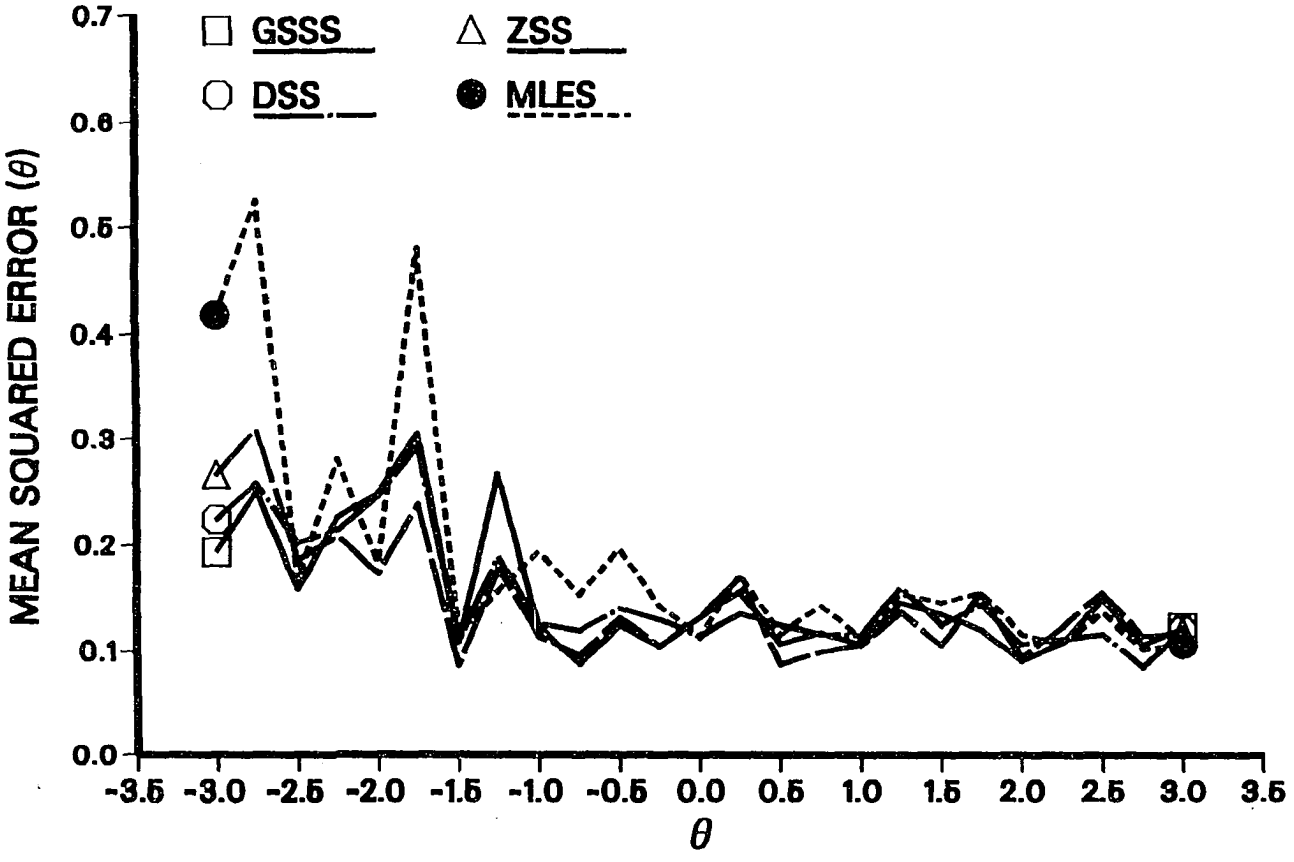

Figure 5.1: Mean squared errors for GSSS, DSS, ZSS, and MLES using the quasi-match m_i to $\hat{\theta}$ item selection, in the hypothetical 3-PL item pool

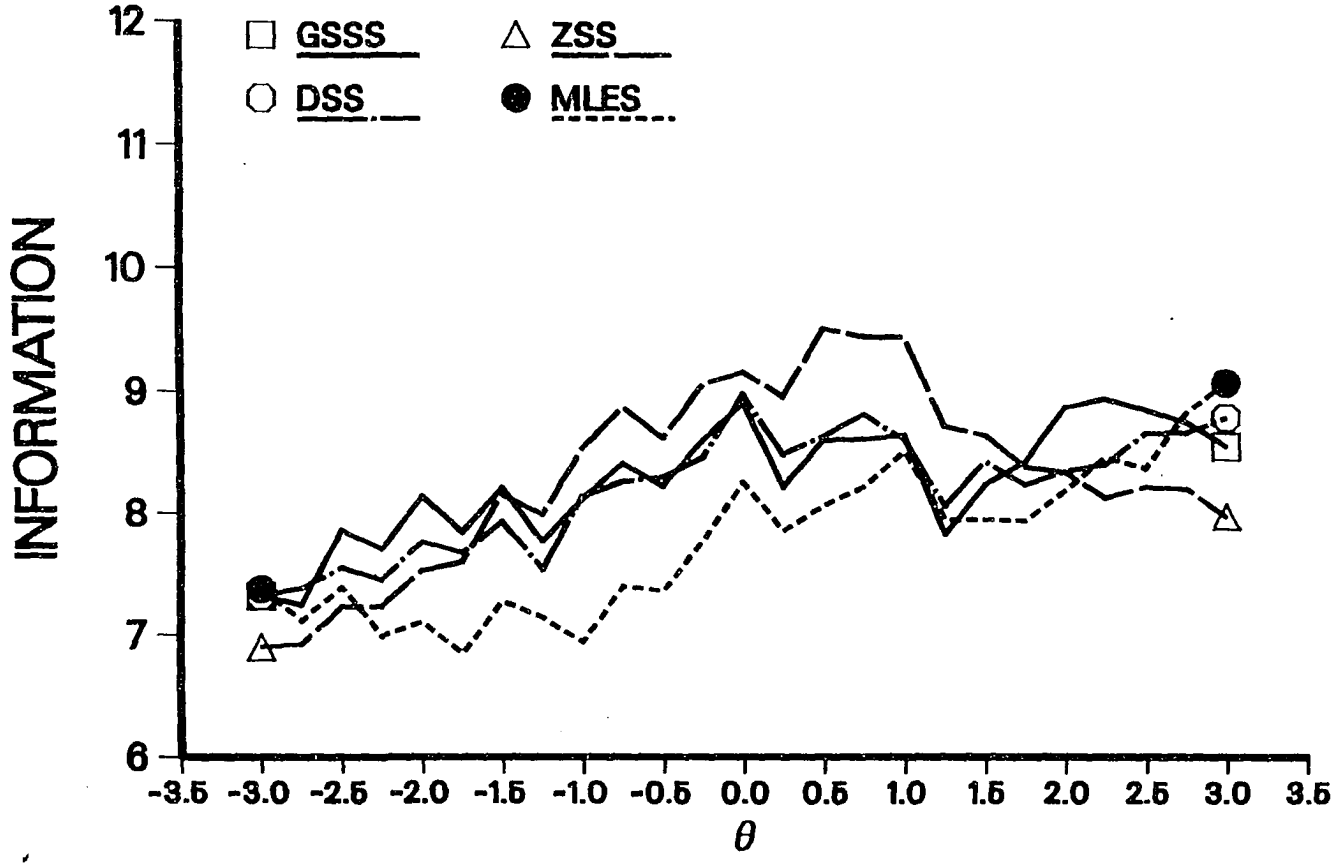

Figure 5.2: Test information for GSSS, DSS, ZSS, and MLES using the quasi-match m_j to $\hat{\theta}$ item selection, in the hypothetical 3-PL item pool

m_i to $\hat{\theta}$ item selection were: 8.3880, 8.3465, 8.3234, and 7.8140, respectively. The four information curves are shown in Figure 5.4. Information obtained by MLES was significantly lower than those of the other CAT strategies along the whole ability continuum, except for extremely lower levels of ability. Information obtained by GSSS, DSS, and ZSS was almost the same. DSS provided slightly higher information than GSSS and ZSS in the extremely higher levels of ability; GSSS and ZSS provided slightly higher information than DSS, in the lower and middle range of ability.

MSE and information using the maximum information item selection. Using the maximum information item selection, the MSEs for GSSS, DSS, ZSS, and MLES were: 0.110, 0.123, 0.123, and 0.140, respectively. Figure 5.5 shows that the MSE for MLES is greater than those for GSSS, DSS, ZSS, especially in the lower range of ability levels.

The means of test information for GSSS, DSS, ZSS, and MLES using the maximum information item selection were: 9.6991, 9.5539, 9.5233, and 9.2061, respectively. Information curves are shown in Figure 5.6. Information obtained by MLES was lower than those of the other CAT strategies, along the whole ability continuum, except for extremely higher or lower levels of ability. Information obtained by GSSS, DSS, and ZSS was almost the same. DSS provided slightly higher information than the other two CAT strategies in the extremely lower or higher levels of ability; GSSS and ZSS provided higher information than DSS in the middle range of ability.

Executing time. Using the ZENITH 386 microcomputers to run the compiled BASIC programs for GSSS, DSS, ZSS, and MLES, the speed was fast. Table 5.7 listed the time needed in GSSS, DSS, ZSS, and MLES to administer 100 simulees, whose

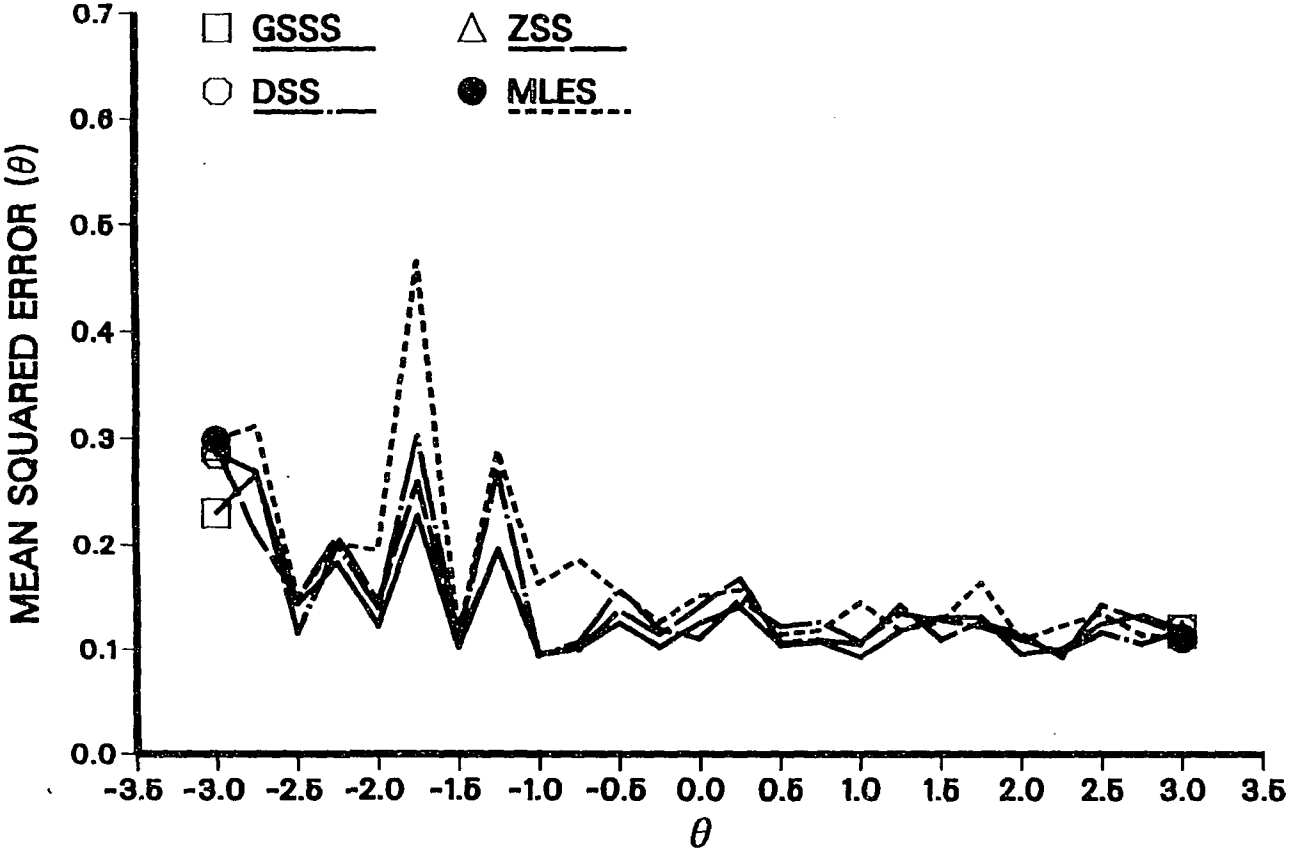

Figure 5.3: Mean squared errors for GSSS, DSS, ZSS, and MLES using the match m_i to $\hat{\theta}$ item selection, in the hypothetical 3-PL item pool

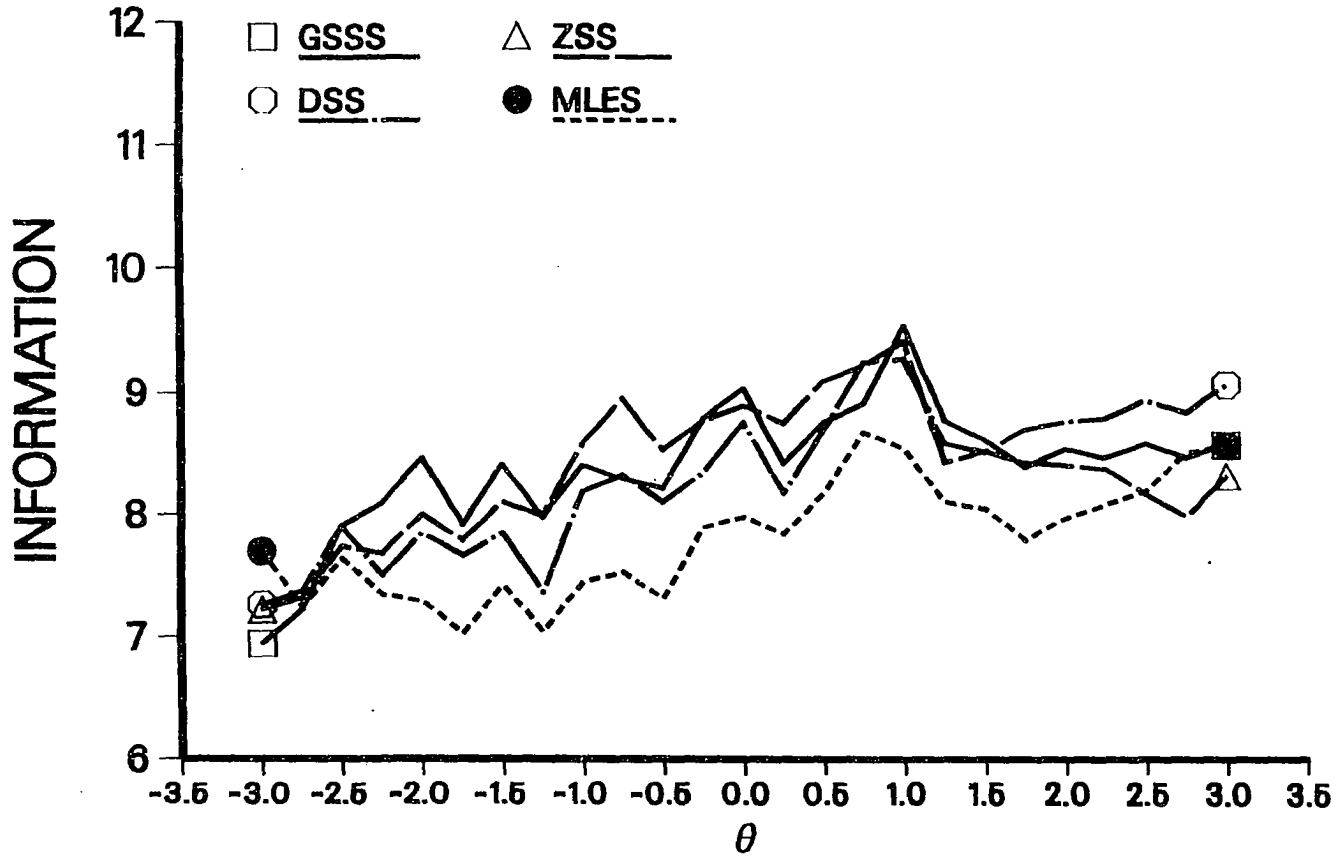

Figure 5.4: Test information for GSSS, DSS, ZSS, and MLES using the match m_j to $\hat{\theta}$ item selection, in the hypothetical 3-PL item pool

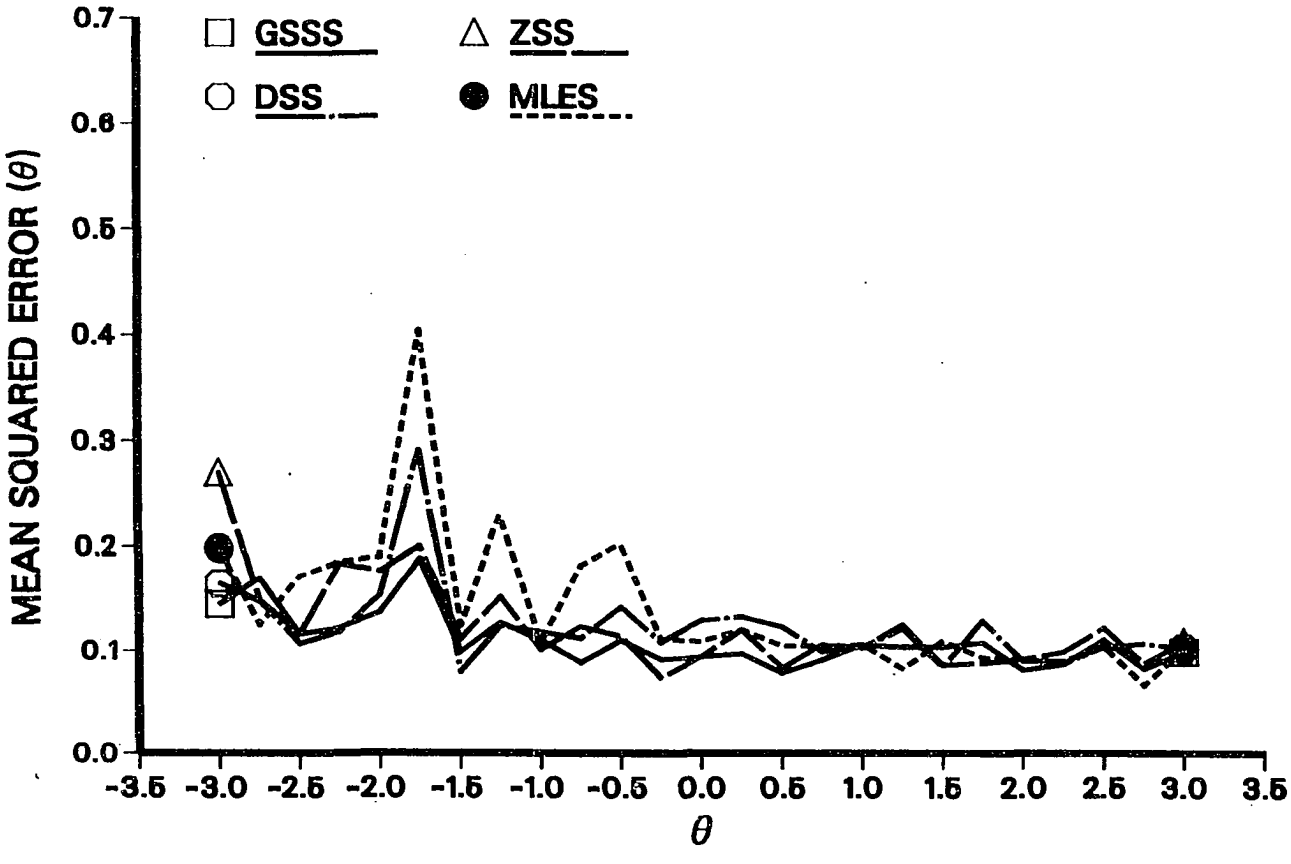

Figure 5.5: Mean squared errors for GSSS, DSS, ZSS, and MLES using the maximum information item selection, in the hypothetical 3-PL item pool

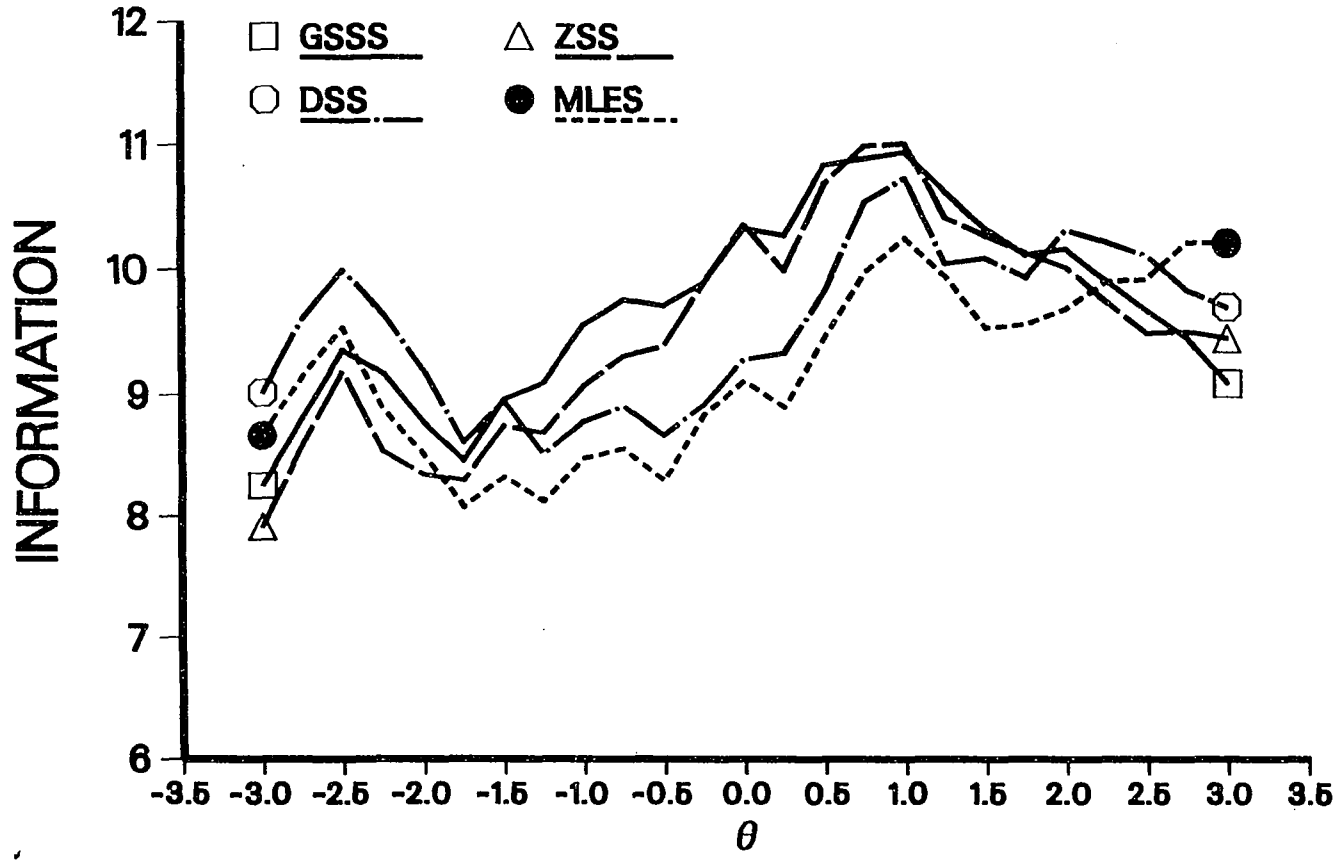

Figure 5.6: Test information for GSSS, DSS, ZSS, and MLES using the maximum information item selection, in the hypothetical 3-PL item pool

ability level was -1.5. Among the four strategies, ZSS was the fastest to operate, while MLES was the slowest to operate. The speed of GSSS and DSS was almost the same. Among the three item selection methods, the quasi-match m_i to $\hat{\theta}$ item selection was the fastest, while the maximum information item selection was the slowest.

Further comparison of the computational efficiency for the four CAT strategies using the the quasi-match m_i to $\hat{\theta}$ item selection and the maximum information item selection was made on the Digital UNIX workstation. The executing time for GSSS, DSS, ZSS (using SD weight), ZSS (no SD weight), and MLES in Study Two or Study Three was listed in Table 5.8. The executing speed of the C programs on the Digital UNIX workstation was about 20 times as fast as that of the BASIC programs on the 386 microcomputer. The speed of using the quasi-match m_i to $\hat{\theta}$ item selection was much faster than that of using the maximum information item selection. Using the maximum information item selection, the executing time for MLES was much more than the executing time for GSSS, DSS, and ZSS. Both versions of ZSS took almost the same amount of time to execute. ZSS was the fastest to operate among the four CAT strategies.

In summary, in the hypothetical 3-PL pool, where there was moderate guessing effect, using the three different item selection methods, the overall measurement accuracy of MLES was significantly worse than those of GSSS and ZSS, was not significantly less accurate than that of DSS; the overall measurement efficiency of MLES was significantly less than those of GSSS, DSS, and ZSS. Using different item selection methods, MLES provided less precise ability estimate than the other CAT strategies. The maximum information item selection method provided the most accurate and efficient ability estimates among the three item selection methods. The

Table 5.7: Executing time for GSSS, DSS, ZSS, and MLES on IBM compatible ZENITH 386/20 PC, in the hypothetical 3-PL item pool[ab]

Method	GSSS	DSS	ZSS	MLES
Quasi-match m_i to $\hat{\theta}$	52"	52"	30"	132"
Match m_i to $\hat{\theta}$	110"	110"	90"	158"
Maximum Information	177"	175"	156"	227"

[a]Compiled BASIC programs. 100 simulees in each cell whose ability levels were -1.5.

[b]SD weight = 0.7 for GSSS, DSS, and ZSS.

match m_i to $\hat{\theta}$ item selection was not significantly more precise than the quasi-match m_i to $\hat{\theta}$ item selection in measurement accuracy, but was in measurement efficiency. Within each item selection method, MLES was the strategy that provided the least precise ability estimates among the four CAT strategies. For the speed of the CAT programs, no matter what item selection method was used, MLES took much more time to operate than did GSSS, DSS, and ZSS. ZSS was the fastest strategy to operate among the four CAT strategies.

Discussion

Results of Study Four showed that in the hypothetical 3-PL item pool, GSSS, DSS, and ZSS provided more precise ability estimates than did MLES. Using any

Table 5.8: Executing time for GSSS, DSS, ZSS, and MLES on Digital UNIX DEC Station 3100, in the modified 3-PL item pool[a]

Method	GSSS	DSS	ZSS[b]	ZSS[c]	MLES
Quasi-Match m_i to $\hat{\theta}$	70"	68"	45"		
Maximum Information	179"	177"	157"	151"	292"

[a]Compiled C programs. 2500 simulees in each cell. SD weight = 0.7 in GSSS and DSS.

[b]ZSS using SD weight = 0.7.

[c]ZSS without SD weight.

of the three item selection methods, MLES took much more time to operate than the other three strategies—GSSS, DSS, and ZSS. ZSS took less time to operate than GSSS and DSS.

"The best and most sophisticated adaptive program cannot function if it is held in check by a limited pool of items, or items of poor quality" (Flaugher, 1990). In Study Four, an hypothetical item pool was used. Though it was not generated from real test situation, its difficulties were widely spread, the average discrimination parameter was moderate and the average guessing parameters was also moderate. Those kinds of items were frequently seen in the literature (e.g, Lord, 1968; Hulin, et al., 1983), except that those items might not have as wide a range of difficulty as the present study, or the values of discrimination parameter might be too small in the extremely lower or higher levels of difficulty range. The present research intended to

examine the measurement quality of several CAT strategies in a broad ability range test situation. Wider difficulty range was chosen, and the probability of getting good quality items was the same for almost every level of ability. In this way, the research results might be more generally applied in any broad range tests.

To generate item responses, the same sequence of random number was used in order to eliminate the errors caused by difference in random number accessing. Though the test condition was identical for every CAT strategy, GSSS, DSS, and DSS measured more accurately and more efficiently than did MLES.

Results of Study Four showed the robust and precise nature of GSSS, DSS, and ZSS using any of the item selection methods. In MLES the current ability estimate was the MLE of ability, and the next item to be chosen was based on the MLE of ability. In GSSS, DSS, and ZSS, the current ability estimate was an approximation of the MLE of ability. It was in the confidence interval of MLE at a testing point which was determined by a certain optimal way. The three item selection methods—the quasi match m_i to $\hat{\theta}$, the match m_i to $\hat{\theta}$, and the maximum information item selections—select the item that is most suitable to the examinee's current ability estimate, according to their own item selection algorithms. Among the three, the maximum information item selection was the most efficient item selection method. It selected from the entire item pool the items not yet administered, the most informative item for that current ability estimate. In the 3-PL item pools, in order to control the item exposure rate or reduce computational burden, a less efficient item selection method, such as the match m_i to $\hat{\theta}$ item selection, or the quasi-match m_i to $\hat{\theta}$, could be used.

Results of Studies Three and Four showed that GSSS, DSS, and ZSS were more

robust against guessing effect, more computationally efficient than MLES, no matter what item selection methods was used.

CHAPTER 6. STUDY FIVE: MEASUREMENT PRECISION FOR GSSS, DSS, ZSS, AND MLES IN THE SAT VERBAL 3-PL POOL ASSUMING THE 3-PL AND THE 1-PL MODELS

In the previous studies, hypothetical item pools were used instead of item pools from real testing. The advantage of using a hypothetical item pool includes that it could be chosen according to particular requirements. The disadvantage of it is the hypothetical item pool might not be compatible to item pools from real testing. Study Five intended to use a real test item pool, to examine the measurement precision of GSSS, DSS, ZSS, and MLES. The present study also intended to test the robustness of CAT strategies against the inaccuracy of item parameters. It is expected that GSSS, DSS, and ZSS should be more robust against the inaccuracy of the item parameters than MLES and provide more accurate and more efficient ability estimates than MLES.

Design

Monte Carlo studies were conducted, to compare measurement precision of GSSS, DSS, ZSS, and MLES, in a SAT Verbal 3-PL item pool, assuming the 3-PL model, and the 1-PL model.

Measurement accuracy and efficiency of ability estimates using different CAT

strategies, assuming different IRT models, were compared. For the four CAT strategies, each CAT contained 20 items. Two estimated precision indices—absolute errors and test information—were used separately as dependent variable in Random Block Factorial (RBF) 4 (CAT Strategy) x 2 (Test Model) ANOVA analysis. Two other estimated precision indices, bias, and MSE in different CAT strategies, were compared. The fidelity correlations—the correlation of the ability estimates and the true ability—were computed for each test strategy assuming each test model. All tests were conducted by Digital UNIX DEC Station 3100 using simulated examinees.

Method

Simulees

A simulee was a computer generated simulee with a true ability value θ. In each of the four CAT strategies (GSSS, DSS, ZSS, and MLES) assuming each of the two IRT models (the 3-PL and the 1-PL), there would be 1000 simulees normally distributed with mean 0 and variance 1. The abilities of the 1000 simulees were generated by a SAS program (see Appendix B).

Item pool

Item parameters of 138 3-PL items had been selected from a 430 SAT Verbal item pool. The 430 items were 5-choice multiple choice items. Item parameters were estimated from a recent SAT administration, calibrated concurrently with LOGIST Version 5, on four operational verbal forms of the test and two sets of external linking (equating) items by the College Board. The equating tests were similar in content and statistical specifications to the operational tests. The examinee samples consisted

of 5500-6000 high school juniors and seniors who took each of the four forms. To compile an item pool for CAT use, the author developed a program that could select the most informative 20 items from the entire 430 item pool for each ability level from -3.5 to 3.5, in interval 0.01. There were 138 non-repeated items selected for the present research. Appendix D contains the 138 SAT Verbal 3-PL item parameters. The range for as was [0.468, 1.643], for bs was [-4.181, 3.390], for cs was [0.000, 0.395], respectively. The mean and standard deviation for as were 1.027, 0.253, for bs were -0.005, 1.802, for cs were 0.149, 0.076, respectively. The correlation between as and bs was 0.427 ($p < 0.001$), between as and cs was 0.466 ($p < 0.001$), between bs and cs was 0.232 ($p < 0.01$).

CAT strategies

Maximum information item selection was used in each CAT strategy. MLE of ability was used in the final ability estimation for each CAT strategy. Every simulee's ability level was assumed to be $\theta = 0$ before testing. In final ability estimation, if a simulee's item responses were all 1s (or all 0s), or a simulee's MLE solution was less than -10 (or greater than 10), it would be excluded from the results. Each excluded case was replaced by a simulee with the same ability level.

In GSSS and DSS, the range of the original search region was [-3.2, 3.2]. The current ability estimate was limited by that range. GSSS, and DSS are the same as in Study Three. SD weight 0.7 was used in determining the confidence interval of the expected score at each testing point. ZSS (no SD weight) that was used was the same as ZSS (no SD weight) in Study Three. The current ability estimate in ZSS

was limited to the range of the item difficulty, which was [-4.181, 3.390]. MLES was the same as in Study Three.

Test models assumed

There were two test models assumed: the 3-PL model and the 1-PL model. The item pool was a 3-PL pool. When assuming the 3-PL model, the item responses were generated according to the 3-PL item parameters. The ability estimation and item selection were also calculated using the 3-PL item parameters. When assuming the 1-PL model, though the item responses were generated according to the 3-PL item parameters, the ability estimation and item selection were calculated using the 1-PL item parameters, in which $a\prime = 1$, $c\prime = 0$ for all items. For item i, the $b\prime$ in the 1-PL model was calculated by the following formula

$$b_i\prime = b_i - \frac{1}{Da_i} \ln \frac{1}{1 - 2c_i} \tag{6.1}$$

where a_i, b_i, and c_i were item parameters of item i in the 3-PL pool.

One can calculate from Equation 1.1 that when the ability

$$\theta = b_i - \frac{1}{Da_i} \ln \frac{1}{1 - 2c_i}$$

the probability of a keyed response to item i for an individual whose ability level is θ is .5. The difficulty value $b_i\prime$ in the 1-PL model was slightly smaller than the corresponding 3-PL item difficult value b_i, if the 3-PL guessing parameter $c_i > 0$. An individual whose ability level was equal to the $b_i\prime$ could get 50% chance to respond to the item correctly.

Results

In the SAT Verbal 3-PL item pool, there was significant difference among the measurement precision of GSSS, DSS, ZSS and MLES. The bias, absolute errors, mean squared errors, test information, and the fidelity correlations—the correlation of the ability estimates and the true ability—were computed for each test strategy assuming each test model.

Bias and frequencies of replacement. Table 6.1 lists the average bias of 1000 normally distributed simulees for GSSS, DSS, ZSS, and MLES, in the SAT Verbal item pool, assuming the 3-PL model, and the 1-PL model. The average bias for any of the CAT strategies was very small. No replacement was made for DSS, ZSS, and MLES. One simulee in GSSS was replaced by another simulee with the same ability level.

Absolute errors. Table 6.1 lists the average absolute errors for GSSS, DSS, ZSS, and MLES assuming each of the two test models. When assuming the 3-PL model, the mean absolute errors for GSSS, DSS, and MLES were almost the same. The mean absolute errors obtained by ZSS was slightly smaller than those obtained by the other three CAT strategies. When assuming the 1-PL model, the smallest mean of absolute errors was obtained by GSSS. The mean absolute errors for DSS and ZSS was slightly bigger than that for GSSS. The mean absolute errors obtained by MLES was much bigger than those obtained by GSSS, DSS, and ZSS. The overall means of absolute errors for GSSS, ZSS, DSS, and MLES were: 0.261, 0.261, 0.271, and 0.293, respectively. Results of the RBF 4 (CAT Strategy) x 2 (Test Model) ANOVA

Table 6.1: Measurement precision for GSSS, DSS, ZSS, and MLES in the SAT Verbal 3-PL item pool[a]

CAT strategy	Bias	ABE	MSE	Info	Replace
Assuming 3-PL					
GSSS	-0.012	0.245	0.094	11.375	1
DSS	-0.007	0.242	0.097	11.189	0
ZSS	-0.002	0.234	0.088	11.428	0
MLES	-0.011	0.246	0.099	10.768	0
Assuming 1-PL					
GSSS	0.029	0.277	0.129	9.234	0
DSS	0.067	0.301	0.174	8.636	0
ZSS	0.030	0.289	0.149	9.114	0
MLES	0.073	0.340	0.230	7.871	0

[a]1000 normally distributed simulees in each cell.

analysis are shown in Table 6.2. Difference among the means of absolute errors for GSSS, ZSS, DSS, and MLES was significant ($F(3,2997) = 15.31$, $p < .01$). Further comparison among the four means using LSD ($\alpha = 0.01$, $df = 2997$, $MSE = 0.02905$, $n = 2000$) showed that the mean absolute errors for MLES was significantly greater than those for GSSS, DSS, and ZSS. The means of absolute errors for GSSS, ZSS, and DSS were not significantly different.

The overall means of absolute errors assuming the 3-PL and the 1-PL models were 0.242 and 0.302, respectively. Difference among the means of absolute errors assuming the two test models was significant ($F(1,999) = 79.61$, $p < .01$). The interaction between the CAT strategies and the test model assumed was also significant

Table 6.2: ANOVA table for RBF 4x2 design in the SAT Verbal 3-PL item pool, with absolute errors as the dependent variable

Source	*SS*	*df*	*MS*	*F*
B (CAT strategy)	1.3344	3	0.4448	15.31*
BS	87.0662	2997	0.0291	
C (Model assumed)	7.2243	1	7.2243	79.61*
CS	90.6549	999	0.0907	
S (Simulee)	200.5586	999	0.2008	7.35*
BC	1.0004	3	0.3335	12.20*
BCS	81.8900	2997	0.0273	
Total	469.7288	7999		

*$p < .01$

($F(3, 2997) = 12.20$, $p < .01$).

MSE. Table 6.1 shows the MSEs for GSSS, DSS, ZSS, and MLLES assuming the 3-PL and the 1-PL models. The MSEs for the four CAT strategies were almost the same when the 3-PL model was assumed. The MSE for MLES was much bigger than those for the other three CAT strategies when the 1-PL model was assumed. Bigger MSEs were obtained when the 1-PL model was assumed than those when the 3-PL model was assumed. Figure 6.1 shows the MSEs for GSSS, DSS, ZSS, and MLES assuming the 3-PL and the 1-PL models. Among the four CAT strategies the differences in MSEs did not occur when assuming the 3-PL model. When the 1-PL

model was assumed, the MSE for MLES was much greater than those for GSSS, DSS, and ZSS. GSSS obtained the smallest MSE among the four CAT strategies. ZSS obtained the second smallest MSE. DSS obtained the third smallest MSE.

Overall test information. Table 6.1 lists the average information for GSSS, DSS, ZSS, and MLES assuming each of the two test models. From Table 6.1 one could see that no matter what test model was assumed, the means of information for GSSS, DSS, and ZSS were higher than that obtained by MLES. The overall means of information for GSSS, ZSS, DSS, and MLES were: 10.304, 10.271, 9.913, and 9.320, respectively. Results of the RBF 4 (CAT Strategy) x 2 (Test Model) ANOVA analysis are shown in Table 6.3. Difference among the means of information for GSSS, ZSS, DSS, and MLES was significant ($F(3, 2997) = 206.99$, $p < .01$). Further comparison among the four means using LSD ($\alpha = 0.01$, $df = 2997$, $MSE = 2.0208$, $n = 2000$) showed that the means of information for GSSS, DSS, and ZSS were significantly higher that that of MLES. Information for GSSS and ZSS was not significantly different. Information obtained by GSSS and ZSS was significantly higher than that of DSS.

The overall means of information assuming the 3-PL and the 1-PL models were 11.190 and 8.714, respectively. Difference among the means of information assuming the two test models was significant ($F(1, 999) = 1453.88$, $p < .01$). The interaction between the CAT strategies and the test model assumed was also significant ($F(3, 2997) = 34.42$, $p < .01$).

Figure 6.2 shows the means of information for GSSS, DSS, ZSS, and MLES assuming the 3-PL and the 1-L models. The information for MLES was lower

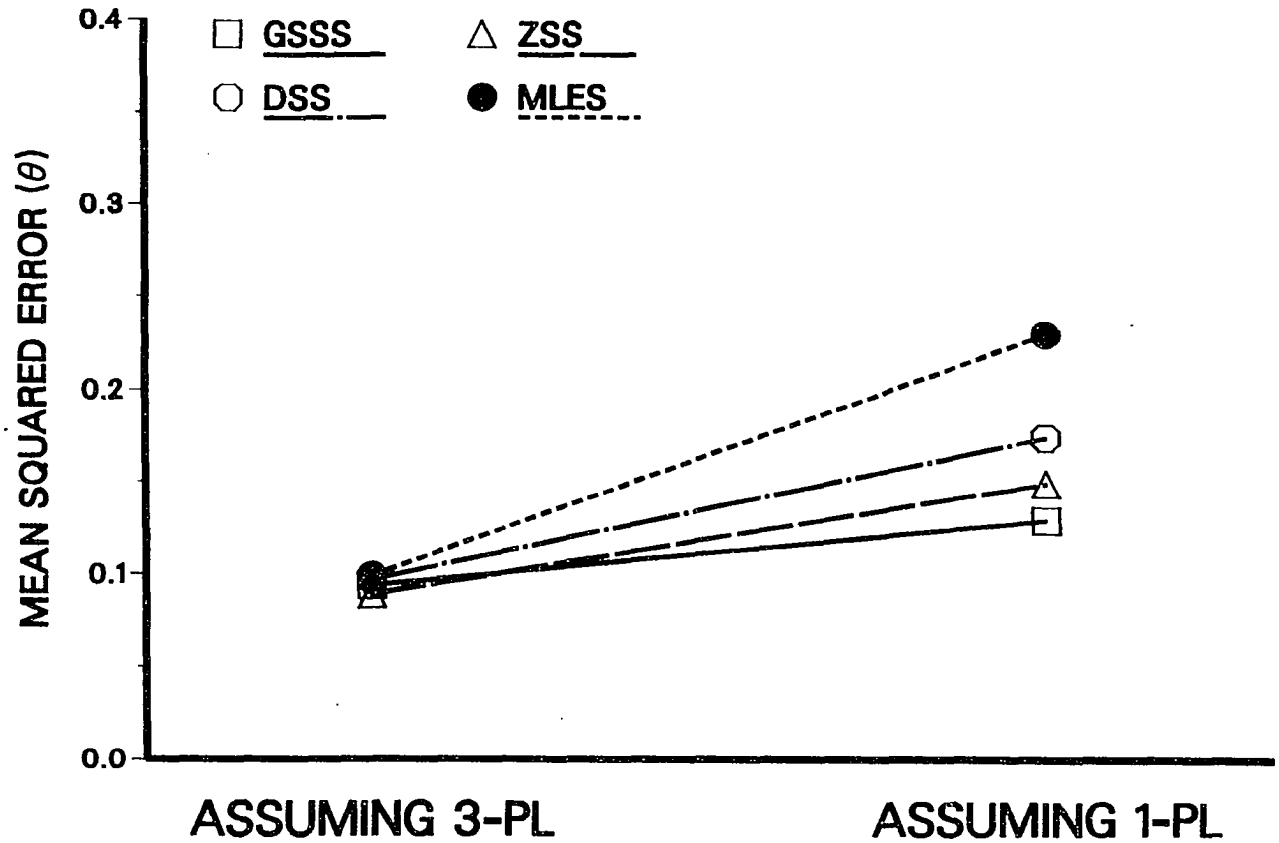

Figure 6.1: Mean squared errors for GSSS, DSS, ZSS, and MLES assuming the 3-PL and the 1-PL models, in the SAT Verbal 3-PL item pool

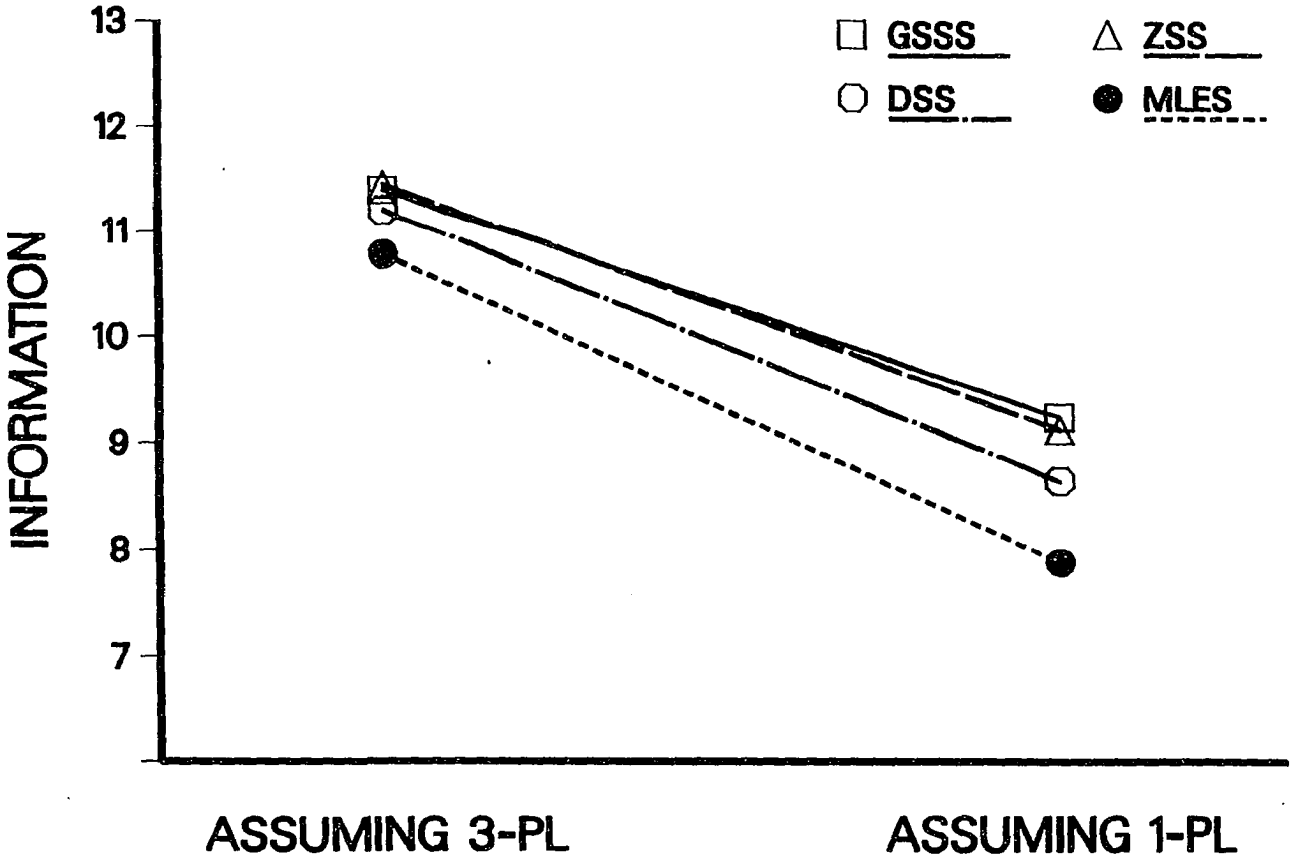

Figure 6.2: Test information for GSSS, DSS, ZSS, and MLES assuming the 3-PL and the 1-PL models, in the SAT Verbal 3-PL item pool

Table 6.3: ANOVA table for RBF 4x2 design in the SAT Verbal 3-PL item pool, with information as the dependent variable

Source	*SS*	*df*	*MS*	*F*
B (CAT strategy)	1254.8840	3	418.2947	206.99*
BS	6056.4505	2997	2.0208	
C (Model assumed)	12260.5726	1	12260.5726	1453.88*
CS	8424.5510	999	8.4330	
S (Simulee)	35519.2342	999	35.5548	22.80*
BC	160.9841	3	53.6614	34.42*
BCS	4672.9633	2997	1.5592	
Total	68349.6397	7999		

*$p < .01$

than those for GSSS, DSS, and ZSS no matter what test model was assumed. The difference between the information for MLES and the other three CAT strategies was much greater when the 1-PL model was assumed than that when the 3-PL model was assumed. The highest information was obtained by GSSS among the four CAT strategies. The second highest information was obtained by ZSS. The third highest information was obtained by DSS. The information obtained assuming the 1-PL model was much lower than that if the 3-PL model was assumed.

Fidelity correlations. Fidelity correlations—the correlation of the ability estimates and the true ability levels for each CAT strategy assuming each test

model—are shown in Table 6.4. The fidelity correlations for GSSS, DSS, and ZSS were higher than that for the MLES. The fidelity correlations obtained assuming the 3-PL model were higher than those obtained assuming the 1-PL model.

Discussion

Study Five used an item pool that was from real testing and a simulated sample of examinees whose ability levels were normally distributed. Results of Study Five suggested that GSSS, DSS, and DSS could measure more accurately and more efficiently than MLES. GSSS, DSS, and ZSS were more robust against inaccuracy of item parameters.

When the 3-PL model was assumed, though, the difference of measurement accuracy between MLES and the other three CAT strategies was very small. The correlations between the estimated abilities and the true ability levels were almost the same for the four CAT strategies. MLES measured less efficiently than did GSSS, DSS, and ZSS. When the item parameter accuracy was deliberately violated (that is, when the 1-PL model was assumed), the measurement precision for MLES was much worse than that for GSSS, DSS, and ZSS. The correlation between the estimated abilities and the true ability levels was lower for MLES than those for GSSS, DSS, and ZSS. The overall measurement accuracy and efficiency for MLES were significantly worse than those for the other three strategies. The reasons for the small difference between measurement precision of MLES and the other three CAT strategies when assuming the 3-PL model might be that the mean of cs for the 3-PL SAT Verbal item pool was moderate (0.149), and the correlation between the cs and the bs was significant. The guessing parameters for the lower difficulty

Table 6.4: Fidelity correlations of true and estimated ability levels for GSSS, DSS, ZSS, and MLES in the SAT Verbal 3-PL item pool[a]

Model assumed	GSSS	DSS	ZSS	MLES
3-PL	.952	.949	.954	.951
1-PL	.932	.912	.924	.895

[a]1000 normally distributed simulees in each cell.

levels were very small. Most of the cs of the lower difficulty items shared a low common c (0.099). GSSS, DSS, ZSS were more robust against the guessing effect than was MLES. The small guessing effect in the lower ability range might result in the small measurement precision difference between MLES and the other three CAT strategies. The correlation between the as and the bs of the SAT Verbal item pool was very strong. It is suggested that the items with lower difficulties usually have lower discrimination power. Lower discrimination power results in lower item information. Lack of informative items in the lower levels of ability might override the advantages of GSSS, DSS, and ZSS, which measured much better than did MLES in the range of lower and middle ability levels.

Previous research has shown that CAT substantially enhances the efficiency of testing. The difficulty to calibrate an appropriate item pool for a CAT might prohibit the wide use of CAT. More parameters are included in a model, more examinees and items are needed to achieve an acceptable level of accuracy of the item parameter estimation. The inaccuracy of item parameter estimation might occur if

the sample size is not big, or the true model should include more parameters than the model assumed. Using more robust CAT strategies, the effect of inaccuracy of item parameters can be drastically reduced.

GSSS, DSS, and ZSS were robust against inaccuracy of item parameters. When the test model assumed was the 3-PL, which was the same as the true item parameters, the measurement accuracy for MLES was not significantly worse than that for GSSS, DSS, and ZSS. However, when the 1-PL model was assumed, there was discrepancy between the true item parameters and the assumed item parameters. The assumed item parameters differed from the true 3-PL parameters to a certain extent. The item responses were generated from the true 3-PL item parameters while the ability estimation and item selection process were based on the assumed 1-PL item parameters. In that situation GSSS, DSS, and ZSS measured much better than did the MLES. GSSS, DSS, and ZSS were more robust against the inaccuracy of item parameters than did MLES.

CHAPTER 7. CONCLUSIONS

The present research introduced three CAT strategies—GSSS, DSS, and ZSS. GSSS, DSS, and ZSS (using SD weight) applied statistical hypothesis testing in determining the current ability estimates. The successive testing points were determined by optimization algorithms in GSSS and DSS, by the Z-score estimation in ZSS (using SD weight). ZSS (no SD weight) used Z-score estimate with respect to the previous ability estimate as the current ability estimate. Results showed that GSSS, DSS, and ZSS strategies measured more precisely and took less computer time to operate than did MLES. GSSS, DSS, and ZSS were more robust against random guessing effect and against inaccuracy of item parameters than MLES.

Study One explored the optimal SD weight for operating GSSS, DSS, and ZSS (using SD weight), in the 1-PL and the modified 3-PL item pools. Results showed that the optimal SD weight was not too big or too small. A SD weight of 0.7 achieved satisfactory measurement quality for GSSS, DSS, and ZSS not only in the 1-PL and the modified 3-PL item pools, but also in the later studies using the 3-PL item pools.

Study Two compared the measurement precision and computational efficiency using the quasi-match m_i to $\hat{\theta}$ item selection and the maximum information item selection methods for GSSS, DSS, and ZSS, and also compared the measurement precision and computational efficiency for ZSS (using SD weight) and ZSS (no SD

weight). The measurement precision for using the two item selection methods was not different when the item was a 1-PL or the modified 3-PL item pool. Using the quasi-match m_i to $\hat{\theta}$ item selection could drastically reduce the executing time in GSSS, DSS, and ZSS. The measurement precision of the two versions of ZSS was also not different in the 1-PL or the modified 3-PL item pools.

Study Three compared the measurement precision of GSSS, DSS, ZSS, and MLES, in the 1-PL and the modified 3-PL item pools. GSSS, DSS, and ZSS measured significantly more efficiently than did MLES in both of the item pools. GSSS, DSS, and ZSS also measured significantly more accurately than did MLES in the modified 3-PL item pool, but not significantly more accurately than did MLES in the 1-PL item pool.

Study Four compared the measurement precision of the four CAT strategies using three item selection methods—the quasi-match m_i to $\hat{\theta}$, the match m_i to $\hat{\theta}$, and the maximum information item selections, in a hypothetical 3-PL item pool. The overall measurement efficiency of GSSS, DSS, or ZSS was significantly higher than that of MLES. The overall measurement accuracy of GSSS or ZSS was significantly more precise than that of MLES. The measurement accuracy of DSS and MLES was not significantly different. The maximum information item selection provided significantly more efficient and more accurate ability estimates than the other two item selection methods. The match m_i to $\hat{\theta}$ item selection was significantly better than the quasi-match m_i to $\hat{\theta}$ item selection in measurement efficiency, but not in measurement accuracy.

Study Five compared the measurement precision of the four CAT strategies in a SAT Verbal 3-PL item pool assuming the 3-PL and the 1-PL models. Unlike

the previous studies in which examinees were grouped into 25 ability levels from -3 to 3 in interval 0.25, examinees whose ability levels were normally distributed with mean 0 and variance 1 were used in Study Five. Results of Study Five showed that GSSS, DSS, and ZSS measured more accurately and more efficiently than did MLES. The difference in measurement precision between MLES and the other three CAT strategies was smaller when the 3-PL item pool was assumed and was much greater when the 1-PL model was assumed, in which the item parameter accuracy was deliberately violated. The measurement precision was affected by the inaccuracy of the item parameters. When the item parameters assumed differed from the true item parameters, the measurement precision of GSSS, DSS, and ZSS was not as severely effected as MLES by the inaccuracy of the parameters.

Being similar in the statistical hypothesis testing in determining the current ability estimate, GSSS, DSS, and ZSS (using SD weight) are different in the successive testing points allocation. ZSS (no SD weight) used the Z-score estimate as the current ability estimate. GSSS, DSS, and ZSS adjusted for the test scores that were lower than scores obtained by guessing, during the item selection process. In the present research, GSSS and ZSS generally provided slightly more precise ability estimate than did DSS. DSS provided slightly more flat information curves than did GSSS and ZSS. Information of GSSS and ZSS was slightly more bell shaped. If the ability levels are widely spread, DSS may achieve approximately equal precision measurement along the ability continuum. GSSS and ZSS generaly had slightly higher information in the middle range of ability. GSSS measured slightly more precisely than did ZSS in the extremely lower and extremely higher levels of ability. Though a CAT strategy with a bell shaped information curve is not optimal, since it measures better in the

middle range of ability than does in the lower and higher levels of ability and does not provide equal precision measurement results for all ability levels, it has its potential to apply to real test situations—few CAT item pools contain item difficulty broader than those of the present research, and few CATs intend to measure wider range of ability than those in the present research. The present research used a fixed number of items as the termination criterion. The measurement precision for the lower or the higher levels of ability could be improved by using a prespecified level of measurement accuracy as the CAT termination criterion. In this way, more items are needed for examinees in the lower or the higher levels of ability.

In general, if a wider range of ability must be estimated, DSS might be the first choice. GSSS might be the candidate for measuring a less broad range of ability. ZSS might be suitable when the CAT is intended to be administered to a group with a less broad range of ability. Among GSSS, DSS, ZSS, and MLES, GSSS seemed to be the most robust against the inaccuracy of item parameters. ZSS is the simplest CAT strategies to operate.

The high quality of measurement achieved by GSSS, DSS, and ZSS in all item pools and the simplicity of their operation, in the broad range ability test situation, suggest their robustness against aberrant responses, against errors of item parameter estimation, and their effectiveness in determining current ability estimate. The three CAT strategies can be easily adapted for mastery testing purpose, to achieve efficient and accurate mastery classifications.

Several researchers (e.g., Green, Bock, Humphreys, & Reckase, 1984) suggested that if the items in a CAT item pool have been highly selected, fixed both a and c for all items may be useful. Study Three examed this situation in the modified

3-PL item pool, in which GSSS, DSS, and ZSS measured significantly more precisely than did MLES. In the present research, there was no guessing effect in the 1-PL item pool, guessing effect was moderate in the hypothetical and the SAT Verbal 3-PL item pool, and was greater in the modified 3-PL item pool. The greater the guessing effect, the greater the measurement precision difference between MLES and the other three CAT strategies. In a 3-PL CAT item pool, if the a_is, and c_is do not differ significantly, the quasi-match m_i to $\hat{\theta}$ and the match m_i to $\hat{\theta}$ item selection can operate very efficiently, and select item much more evenly than the maximum information item selection.

The present research found that using a moderate SD weight in GSSS, DSS, and ZSS (using SD weight) could provide very precise ability estimate. Using any SD weight in the range (0.5, 0.9) also might provide optimal measurement quality, depending upon the nature of the item pool and the ability distribution of the examinee group. In GSSS, DSS, and ZSS (using SD weight), SD weight is used for determining the size of a confidence interval of the expected score at a testing point. At the earlier stage of a CAT, the uncertainty of the examinee's ability estimate is greater. As the CAT continues, more items are administered and the uncertainty of the examinee's ability estimate is reduced. It may be more optimal to use SD weight in GSSS, DSS, and ZSS, in which SD weight is a function of the degree of uncertainty of the current ability estimate. The more the degree of uncertainty, the greater the SD weight. The index for that uncertainty could be the reciprocal of test information evaluated at the current ability estimate, or, approximated by the reciprocal of the number of items administered. The smaller the SD weight, the closer the current ability estimate approaches to the MLE. The suggestion of using a SD weight that

decreases accordingly with the uncertainty of the ability estimate is only necessary when item pool is very, very big. If an item pool is not very big, using a constant SD weight can still choose the most informative item that is not yet administered, since items that can provide higher information to a certain ability level are widely distributed, if the item pool is not big enough. This is the case in most practical test situations.

The present research was a simulated one. Care must be taken to apply the results to real test situation. Further studies are need to apply GSSS, DSS, and ZSS to real test situation.

REFERENCES

Adby, P. R., & Dempster, M. A. H. (1974). *Introduction to optimization methods.* Landon: Chapman and Hall.

Angoff, W. H., & Huddleston, E. M. (1958). *The multi-level experiment:A study of a two-level test system for the College Board Scholastic Aptitude Test* (Statistical Report SR-58-21). Princeton, NJ: Educational Testing Service.

Betz, N. E., & Weiss, D. J. (1973). *An empirical study of computer-administered two-stage ability testing* (Research Report 73-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Betz, N. E., & Weiss, D. J. (1974). *Simulation studies of two-stage ability testing* (Research Report 74-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (Part 5, pp. 397-479). Reading, MA: Addison-Wesley.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 413-444.

Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1989). The four generations of computerized educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 367-408). New York: American Council on Education.

Flaugher, R. (1990). Item pools. In Wainer, H., *Computerized adaptive testing: A primer* (pp. 41-63). Hillsdale, NJ: Lawrence Erlbaum.

Gottfried, B. S. & Weisman, J. (1973). *Introduction to optimization theory.*

Englewood Cliffs, NJ: Prentice-Hall.

Green, B. F. (1983). The promise of tailored tests. In H. Wainer, & S. Messick (Eds.), *Principles of modern psychological measurement* (pp. 69-80). Hillsdale, NJ: Lawrence Erlbaum.

Green, B. F., Bock, R. D., Humphreys, L. G., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*, 347-360.

Hamming, R. W. (1973). *Numerical methods for scientists and engineers* (2nd ed.). New York: McGraw-Hull.

Hua, Luogen (1981). 优选法 (*Optimization methods*). Beijing: Science Press.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement.* Homewood, IL: Dow Jones-Irwin.

Jensema, C. J. (1974). An application of latent trait mental test theory. *British Journal of Mathematical and Statistical Psychology, 27*, 29-48.

Krathwohl, D. R., & Huyser, R. J. (1956). The sequential item test (SIT). *American Psychologist, 2*, 419.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monograph.* No. 7.

Lord, F. M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28*, 989-1020.

Lord, F. M. (1970). Some test theory for tailored testing. In W. H. Holtzman (Ed.), *Computer-assisted instruction, testing, and guidance* (pp. 139-183). New York: Harper and Row.

Lord, F. M. (1971a). The self-scoring flexilevel test. *Journal of Educational Measurement, 8*, 147-151.

Lord, F. M. (1971b). A theoretical study of the measurement effectiveness of flexilevel tests. *Educational and Psychological Measurement, 31*, 805-813.

Lord, F. M. (1971c). Robbin-Monro procedure for tailored testing. *Educational and Psychological Measurement, 31*, 3-31.

Lord, F. M. (1980). *Application of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

McBride, J. R. (1975). Problem: Scoring adaptive tests. In D. J. Weiss (Ed.), *Computerized adaptive trait measurement: Problems and prospects* (Research Report 75-5). Minneapolis: University of Minnesota. Department of Psychology, Psychometric Methods Program.

McBride, J. R. (1986). *A computerized adaptive edition of the differential aptitude tests.* Paper presented at the annual meeting of the American Psychological Association, Washington, DC.

Moreno, K. E., Weztel, C. D, McBride, J. R, & Weiss, D. J. (1984). Relationship between corresponding Armed Services Vocational Aptitude Battery (ASVAB) and computerized adaptive testing (CAT) subtests. *Applied Psychological Measurement, 8*, 155-164.

Norden, R. H. (1973). A survey of maximum likelihood estimation, Part 2. *International Statistical Review, 41*, 39-58.

Owen, R. J. (1969). *A Bayesian approach to tailored testing* (Research Report 69-92). Princeton, NJ: Educational Testing Service.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70,* 351-356.

Parker, S. P. (Ed.). (1984). *McGraw-Hill Dictionary of Science and Technical Terms* (3rd ed.). New York: McGraw-Hill.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *Annals of Statistics, 22*, 400-407.

Roid, G. H. (1986). Computer technology in testing. In B. S. Plake & J. C. Witt (Eds.), *The future of testing* (pp. 29-69). Hillsdale, HJ: Lawrence Erlbaum.

Samejima, F. (1973). A comment on Birnbaum's three-parameter logistic model in the latent trait theory. *Psychometrika, 38,* 221-233.

Samejima, F. (1983). Some methods and approaches of estimating the operating characteristics of discrete item responses. In H. Wainer, & S. Messick (Eds.),

Principles of psychological measurement (pp. 159-182). Hillsdale, NJ: Lawrence Erlbaum.

Stocking, M. L. (1987). Two simulated feasibility studies in computerized adaptive testing. *Applied Psychology: An International Review, 36*, 263-277.

Sympson, J. B. (1977). Estimation of latent trait status in adaptive testing procedures. In D. J. Weiss (Ed.), *Applications of computerized adaptive testing* (Research Report 77-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Method Program.

Thissen, D., & Mislevy, R. J. (1990). Testing Algorithms. In Wainer, H., *Computerized adaptive testing: A primer* (pp. 103-135). Hillsdale, NJ: Lawrence Erlbaum.

Urry, V. W. (1977). Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement, 14*, 181-196.

Vale, C. D. (1975). Problem: Strategies of branching through an item pool. In D. J. Weiss (Ed.), *Computerized adaptive trait measurement: Problems and prospects* (research Report 75-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Wagner, H. M. (1975). *Principles of operational research* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Wainer, H., & Mislevy R. J. (1990). Item response theory, item calibration and proficiency estimation. In Wainer, H., *Computerized adaptive testing: A primer* (pp. 65-102). Hillsdale, NJ: Lawrence Erlbaum.

Walsh, G. R. (1975). *Methods of optimization.* London: John Wiley and Sons.

Weiss, D. J. (1973). *The stratified adaptive computerized ability test* (Research Report 73-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Weiss, D. J. (1974). *Strategies of adaptive ability measurement* (Research Report 74-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Weiss, D. J. (1976). *Adaptive testing research at Minnesota-Overview, recent*

research and future directions. Procedings of the First Conference on Computerized Adaptive Testing. Washington, D. C.: Personnel Research and Development Center, U.S. Civil Service Commission.

Weiss, D. J. (1979). Computerized adaptive achievement testing. In H. F. O'Neil, Jr. (Ed.), *Proceedings for instructional system development* (pp. 129-164). New York: Academic Press.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Appllied Psychological Measurement, 6*, 473-492.

Weiss, D. J. (1983). Introduction. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 1-8). New York: Academic.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of compuerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361-375.

Weiss, D. J., & McBride, J. R. (1984). Bias and information of Bayesian adaptive Testing. *Applied Psychological Measurement, 8*, 273-285.

Weiss, D. J., & Vale, C. D. (1987). Adaptive testing. *Applied Psychology: An International Review, 36*, 249-262.

Xiao, Beiling. (1989, March). *Golden section search strategies for computerized adaptive testing.* Paper presented at the Fifth International Objective Measurement Workshop, Berkeley, CA.

Xiao, Beiling. (1990, April). *Dichotomous search strategies for computerized adaptive testing.* Paper presented at the American Educational research Association Annual Meeting, Boston, MA.

Yen, W. M., Burket G. R., & Sykes, R. C. (1991). Nonunique solutions to the likelihood equation for the three-parameter logistic model. *Psychometrika, 56*, 39-54.

APPENDIX A: SAS PROGRAM FOR GENERATING THE HYPOTHETICAL 3-PL ITEM POOL

```
DATA ITEM;
INPUT SEED1 SEED2 SEED3;
DO I = 1 TO 200;
CALL RANNOR(SEED1,A1);
A = 1.0 + 0.2*A1; CALL RANUNI(SEED2,B1);
B = 6.8*B1 - 3.4; CALL RANNOR(SEED3,C1);
C= .20 + .03*C1;
OUTPUT;
END;
CARDS;
26753 54981 779397
DATA ITEM2;
SET ITEM;
KEEP A B C;
IF A < 0.8 THES DELETE;
CARDS;
PROC SORT;BY B;
PROC PRINT;VAR A B C;
PROC MEANS;VAR A B C;
/*
```

APPENDIX B: SAS PROGRAM FOR GENERATING THE NORMALLY DISTRIBUTED ABILITY LEVELS

```
DATA NORMAL;
SEED=3265834;
DO I = 1 TO 1000;
ABILITY = RANNOR(SEED);
OUTPUT;
END;
CARDS;
PROC SORT;BY ABILITY;
PROC PRINT;
PROC MEANS;VAR ABILITY;
/*
```

APPENDIX C: ITEM PARAMETERS OF 167 ITEMS IN THE HYPOTHETICAL 3-PL ITEM POOL

Order	a	b	c
001	1.03505	-3.3550	0.220583
002	1.05103	-3.2633	0.170654
003	0.82562	-3.2602	0.193210
004	1.05707	-3.2549	0.197468
005	1.18103	-3.2453	0.174988
006	0.87371	-3.1853	0.175658
007	1.11584	-3.1644	0.193122
008	1.22951	-3.1600	0.220786
009	1.37939	-3.1496	0.180539
010	0.82015	-3.1371	0.182394
011	0.94704	-3.1360	0.202778
012	1.00015	-3.1027	0.222255
013	1.29790	-3.0809	0.251022
014	0.83143	-3.0630	0.231324
015	1.15609	-3.0456	0.232867
016	0.81911	-3.0286	0.176868
017	0.98822	-2.8760	0.201387
018	0.85379	-2.8582	0.232685
019	1.22977	-2.7429	0.248742
020	0.91965	-2.6177	0.166442
021	0.90862	-2.6158	0.198631
022	0.88255	-2.5759	0.162963
023	1.30328	-2.5660	0.201565
024	1.00941	-2.5247	0.239812
025	1.03922	-2.4857	0.204491
026	0.85920	-2.4654	0.178693
027	1.06271	-2.4135	0.255902
028	0.98068	-2.3490	0.178719
029	1.08705	-2.3446	0.227490
030	1.52473	-2.3212	0.182837
031	1.09353	-2.3174	0.147211
032	0.97847	-2.2638	0.182676
033	1.38022	-2.2565	0.189433
034	1.17721	-2.2200	0.188518
035	1.20808	-2.2013	0.165360

Order	a	b	c
036	1.01387	-2.1765	0.198705
037	0.80508	-1.9311	0.186151
038	1.06846	-1.9191	0.194732
039	1.09673	-1.9113	0.240297
040	0.92500	-1.8429	0.242675
041	1.07471	-1.7849	0.207484
042	1.07839	-1.7765	0.201601
043	1.15408	-1.7345	0.231717
044	0.92352	-1.7023	0.156306
045	0.99776	-1.6157	0.183188
046	1.09449	-1.6023	0.173087
047	0.91395	-1.5361	0.204868
048	1.00512	-1.5032	0.225424
049	1.05357	-1.4698	0.228633
050	0.88875	-1.4255	0.179393
051	1.06111	-1.3971	0.158563
052	1.23222	-1.3417	0.206296
053	1.10329	-1.2943	0.211824
054	1.06975	-1.2877	0.172964
055	1.23328	-1.2594	0.222715
056	1.00948	-1.2117	0.215866
057	1.04530	-1.1227	0.195172
058	1.29350	-1.1018	0.215838
059	1.04085	-1.0619	0.136835
060	0.92994	-1.0233	0.249585
061	1.32117	-0.9405	0.203209
062	1.22763	-0.9343	0.194263
063	1.00548	-0.9205	0.180517
064	0.93893	-0.9189	0.197025
065	1.08216	-0.7941	0.216851
066	1.01323	-0.7573	0.212387
067	0.89071	-0.7554	0.199845
068	1.23555	-0.7218	0.198615
069	0.81193	-0.6899	0.177001
070	0.97799	-0.6853	0.180505

Order	a	b	c
071	0.91412	-0.6720	0.140392
072	0.82602	-0.6347	0.192564
073	0.80299	-0.6112	0.178943
074	1.08912	-0.5688	0.140641
075	1.09177	-0.5015	0.227297
076	1.06435	-0.4961	0.204643
077	0.93897	-0.4390	0.217100
078	0.91597	-0.3558	0.184039
079	1.16830	-0.2671	0.191802
080	0.99877	-0.2123	0.151196
081	1.28414	-0.1913	0.207167
082	1.12614	-0.1863	0.191122
083	1.29858	-0.1640	0.193451
084	1.08127	-0.1229	0.207638
085	1.17258	-0.0904	0.142189
086	1.10603	-0.0307	0.207456
087	1.13947	0.0097	0.142665
088	1.17289	0.0605	0.234396
089	0.90248	0.1074	0.180890
090	1.10519	0.1989	0.182128
091	1.11198	0.2380	0.181885
092	0.91516	0.2578	0.192353
093	1.16510	0.2756	0.170317
094	0.85673	0.3387	0.237956
095	0.83019	0.3471	0.133603
096	0.83752	0.3771	0.216701
097	1.14986	0.4029	0.193225
098	1.09532	0.4633	0.195418
099	1.14948	0.4668	0.194026
100	0.93504	0.4899	0.154996
101	1.43514	0.7001	0.229914
102	0.97935	0.7433	0.181898
103	1.19463	0.7472	0.158089
104	1.13813	0.8738	0.220152
105	1.04202	0.8884	0.191804

Order	a	b	c
106	1.01107	0.8925	0.191985
107	1.24317	0.8931	0.205894
108	1.41611	0.9009	0.202859
109	1.13413	0.9026	0.218342
110	1.40183	0.9043	0.155730
111	0.98343	0.9213	0.166637
112	0.95900	1.0387	0.193557
113	0.84179	1.0463	0.199720
114	1.30522	1.0947	0.181934
115	0.97338	1.1068	0.215887
116	1.02776	1.1263	0.210144
117	1.34913	1.4158	0.199766
118	1.14089	1.4501	0.172070
119	0.96059	1.5170	0.186217
120	0.85881	1.5324	0.216763
121	1.28137	1.5973	0.222411
122	1.16913	1.6332	0.229994
123	0.93248	1.6476	0.230604
124	0.81175	1.6735	0.212668
125	1.22599	1.6819	0.149448
126	0.88364	1.6916	0.191881
127	1.10914	1.6926	0.172083
128	0.96955	1.7318	0.162623
129	1.00199	1.7405	0.225400
130	1.04060	1.8259	0.190644
131	1.39400	1.8398	0.234824
132	1.06477	1.8474	0.244236
133	0.90574	1.8530	0.190013
134	1.04085	1.8798	0.192910
135	0.80785	1.8917	0.168422
136	1.11873	1.9499	0.198838
137	1.34739	2.1474	0.197920
138	1.04586	2.2789	0.158106
139	1.26696	2.2982	0.229345
140	1.18551	2.3547	0.173860

Order	a	b	c
141	1.09123	2.4370	0.206720
142	1.24695	2.4567	0.189216
143	0.83968	2.4730	0.233835
144	1.25860	2.5515	0.148137
145	0.86615	2.6178	0.206052
146	1.00973	2.6899	0.194270
147	1.01476	2.6932	0.190589
148	0.99635	2.7034	0.208607
149	0.92190	2.7139	0.246467
150	1.30503	2.7289	0.191025
151	1.14722	2.7474	0.188791
152	0.84605	2.8213	0.167440
153	1.15783	2.8450	0.146626
154	1.02364	2.8822	0.209307
155	1.00730	2.8910	0.207439
156	1.15854	3.0098	0.229477
157	1.11573	3.0563	0.265189
158	0.81465	3.0689	0.220234
159	1.03119	3.0834	0.176442
160	0.87081	3.0894	0.230744
161	0.89204	3.1010	0.192909
162	1.02914	3.2134	0.234311
163	1.21554	3.2746	0.262857
164	1.34624	3.3437	0.226972
165	1.11685	3.3700	0.194346
166	0.82504	3.3791	0.240373
167	1.33419	3.3813	0.186959

APPENDIX D: ITEM PARAMETERS OF 138 ITEMS IN THE SAT VERBAL 3-PL ITEM POOL

Order	a	b	c
001	0.55246	-4.18058	0.09949
002	0.52663	-3.92650	0.09949
003	0.63805	-3.43082	0.09949
004	0.48528	-3.11000	0.09949
005	0.65709	-2.89926	0.09949
006	0.46804	-2.88262	0.09949
007	0.61930	-2.80981	0.09949
008	0.76553	-2.80619	0.09949
009	0.52973	-2.70033	0.09949
010	1.13023	-2.62967	0.09949
011	0.69031	-2.52182	0.09949
012	0.71648	-2.47330	0.09949
013	0.66357	-2.45082	0.09949
014	0.58477	-2.44037	0.09949
015	0.55168	-2.39258	0.09949
016	0.78954	-2.37408	0.09949
017	0.82868	-2.27626	0.09949
018	0.72333	-2.26962	0.09949
019	0.77689	-2.16641	0.09949
020	0.66410	-2.14042	0.09949
021	0.74568	-2.09903	0.09949
022	0.81540	-2.09759	0.09949
023	0.76100	-2.05054	0.09949
024	0.77746	-2.04047	0.09949
025	0.76279	-2.02350	0.09949
026	0.82001	-1.98633	0.09949
027	0.80208	-1.97371	0.09949
028	1.13935	-1.89356	0.06723
029	0.78149	-1.86626	0.09949
030	0.87555	-1.84260	0.03258
031	0.85319	-1.80986	0.09949
032	0.98832	-1.74593	0.00430
033	1.18894	-1.60888	0.18959
034	1.01178	-1.56463	0.05277
035	1.28809	-1.55024	0.17256

Order	a	b	c
036	0.81907	-1.50928	0.00529
037	1.21849	-1.45831	0.19885
038	1.25965	-1.36985	0.27253
039	0.81484	-1.36623	0.02377
040	1.07404	-1.19885	0.31167
041	0.86795	-1.19520	0.05856
042	1.10549	-1.17358	0.25320
043	1.18276	-1.14419	0.16192
044	0.98564	-1.08114	0.20366
045	0.90798	-1.07982	0.13250
046	1.10427	-1.07086	0.30249
047	0.91333	-1.06386	0.00000
048	1.04002	-0.98823	0.17624
049	1.18794	-0.93114	0.25793
050	1.18572	-0.88522	0.14909
051	1.34704	-0.84171	0.12560
052	1.04739	-0.78357	0.10431
053	1.35521	-0.64166	0.19169
054	1.06880	-0.61722	0.31479
055	1.15414	-0.55091	0.37714
056	1.00732	-0.51857	0.15552
057	1.38751	-0.51418	0.20527
058	1.36253	-0.45419	0.27532
059	1.00039	-0.44247	0.13999
060	1.20096	-0.43642	0.18527
061	1.01847	-0.40431	0.18366
062	0.89614	-0.39709	0.06231
063	1.03514	-0.38432	0.22502
064	1.24922	-0.37415	0.23083
065	1.44491	-0.13486	0.39453
066	1.39682	-0.01591	0.31197
067	1.15739	0.01419	0.18879
068	1.05686	0.01822	0.23752
069	1.03647	0.16542	0.05074
070	1.13824	0.19072	0.24725

Order	a	b	c
071	1.22594	0.19940	0.28181
072	1.16726	0.21330	0.13463
073	1.08313	0.28425	0.11132
074	0.93511	0.31970	0.02855
075	1.22554	0.38563	0.06057
076	1.11019	0.40269	0.18079
077	1.02448	0.46278	0.09010
078	1.06378	0.53895	0.04600
079	1.07154	0.56662	0.20750
080	1.27397	0.56711	0.16800
081	1.16264	0.62131	0.21638
082	1.05731	0.72987	0.11336
083	1.09275	0.73006	0.14451
084	0.93842	0.87626	0.03712
085	1.29258	0.88557	0.15112
086	1.23528	0.91883	0.07195
087	1.16536	0.95717	0.24577
088	1.29403	0.97917	0.25614
089	1.48970	1.02461	0.05215
090	1.30830	1.02779	0.15410
091	1.64384	1.03957	0.18533
092	1.08189	1.06556	0.13403
093	1.19170	1.09516	0.05951
094	1.25370	1.16123	0.23565
095	1.05788	1.23077	0.16453
096	1.34879	1.23206	0.31339
097	1.32631	1.26804	0.16949
098	1.30882	1.29679	0.04195
099	1.09490	1.30671	0.14779
100	1.21400	1.34105	0.12084
101	1.24403	1.40251	0.27454
102	1.28954	1.49754	0.15483
103	1.24130	1.60347	0.15715
104	1.09181	1.62447	0.21033
105	1.17923	1.64656	0.17105

Order	a	b	c
106	1.26983	1.70008	0.19709
107	1.17356	1.71356	0.14296
108	1.56233	1.72710	0.12067
109	1.14149	1.79386	0.21974
110	1.07807	1.84216	0.13980
111	1.01796	1.88109	0.18375
112	1.28316	1.88521	0.14854
113	1.58837	1.91318	0.13858
114	1.08948	1.92407	0.21485
115	1.17282	1.98458	0.22959
116	1.02510	2.00566	0.22479
117	1.00385	2.10480	0.17200
118	0.77828	2.10699	0.12089
119	0.96932	2.12270	0.15872
120	1.30564	2.14514	0.16953
121	0.99964	2.16286	0.24058
122	1.15763	2.19345	0.13396
123	1.37451	2.23833	0.15225
124	1.45865	2.25855	0.21337
125	0.92044	2.33742	0.11301
126	0.90550	2.36793	0.10646
127	0.81521	2.37769	0.15673
128	0.85848	2.37924	0.23678
129	0.94710	2.38132	0.17141
130	0.86779	2.40305	0.07805
131	1.15314	2.42055	0.16232
132	0.61311	2.53156	0.06692
133	0.70636	2.54994	0.16976
134	0.96763	2.81444	0.17676
135	0.64599	2.87124	0.07436
136	0.58520	2.92633	0.12473
137	0.82295	3.09117	0.11835
138	0.67101	3.38999	0.12907